AMERICAN JOURNAL
of LAW *and* EQUALITY

AUTOMATION AND ACCESS TO JUSTICE

Colin Doyle*

The United States faces an access-to-justice crisis. Legal technology has emerged as a potential solution, with generative AI offering new possibilities for automating access to justice. But questions linger about whether AI can automate legal reasoning and whether such capabilities would expand access to justice or deepen inequities. Automation would seem to require an AI that possesses a "general" legal reasoning skill: the capacity to apply law to facts across different domains. But generative AI presents a conundrum. LLM capabilities form a "jagged technological frontier," with LLMs excelling at some tasks while failing at others of similar or lesser complexity. LLMs lack general legal reasoning skills, and their outputs are unreliable for tasks that require accuracy and consistency. This might suggest that current AI is incapable of automating legal reasoning work. This Article reframes our conception of legal automation. Because LLM abilities form a jagged frontier, the frontier of potential automation is also jagged. Tasks within the frontier of LLM capabilities are ripe for automation, while tasks at the border or beyond are not. The challenge is identifying which tasks fall where. Because LLMs lack general skills, an LLM's success with one legal task doesn't predict its performance on similar tasks—it may succeed in applying the law to one set of facts and inexplicably fail with an almost identical set of facts. Creating reliable legal automation therefore requires extensive, careful testing and redesign. Because these evaluations are specific to individual legal domains and tasks, the jagged frontier of legal automation will become increasingly jagged based upon where investments are made. Investment in open-source research focused on low-income litigants'

*Associate Professor of Law, Loyola Law School, Los Angeles. I am grateful for helpful input from Jon Choi, Tristin Green, Justin Hughes, Ariel Jurow Kleinman, Daniel Klerman, Mark Koussa, Julia Mendoza, John Nockleby, Michael Serota, Yan Slavinskiy, Aaron Tucker, Dan Traficonte, Lauren Willis, and Adam Zimmerman; for thoughtful comments from participants at the 1st Workshop on Generative AI and Law (GenLaw '24), MAX-Autonomous Systems Policy Institute Workshop at Syracuse University, 3rd and 4th ACM Computer Science and Law Symposium, and USC's Center for Law and Social Science workshop; and for helpful edits from Kieran Murphy.

© 2025 Colin Doyle. Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (CC BY-NC-ND).
https://doi.org/10.1162/AJLE.a.6

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

legal issues can expand access to justice, while investment in closed, proprietary tools risks exacerbating inequality.

INTRODUCTION

The United States faces an access-to-justice crisis. Millions of people confront evictions, debt collections, and family disputes without professional legal help. The consequences devastate lives.[1] Without legal assistance, people lose their homes, suffer domestic abuse, sink into debt, lose custody of their children, and go without necessary medical care.[2] Legal aid organizations can't bridge this gap because they're chronically underfunded. They are forced to turn away potential clients and can resolve only a fraction of the cases that they do accept.[3] The straightforward solution—more funding for more lawyers for more low-income Americans—remains politically unviable.

Courts and legal aid organizations have turned to technology instead. These technologies aim to increase access to the legal system by facilitating the litigation process and providing self-help tools for unrepresented litigants.[4] Self-help tools can equip litigants with the knowledge necessary to navigate the legal system.[5] The complexity of self-help tools varies, ranging from simple pamphlets to hefty reports to interactive websites. More-sophisticated software can take information about a user's situation and provide tailored legal information or guidance. But developing these automated systems requires formalizing law into code. This is a massive undertaking requiring legal and technical experts to work together to manually code complex decision trees capturing every possible application of a legal rule.[6] Constructing these systems has been feasible only within mechanical, calculator-like pockets of law.[7]

Generative AI seems to open up new opportunities for legal automation. Computers doing legal reasoning used to be the stuff of science fiction, but advances in artificial intelligence—particularly large language models (LLMs) like ChatGPT—have catapulted the automation of legal reasoning from niche preoccupation to profession-wide concern.[8]

1 *See* Margaret Hagan, *The Supply and Demand of Legal Help on the Internet, in* Legal Tech and the Future of Civil Justice 199 (David Freeman Engstrom ed., 2023).

2 *Id.*

3 Legal Servs. Corp., The Justice Gap: The Unmet Civil Legal Needs of Low-income Americans 9 (2022).

4 Edwin O. Eboigbe, AI in Legal Analytics: Balancing Efficiency, Accuracy, and Ethics in Contract and Predictive Analysis 25 (Oct. 7, 2024) (unpublished manuscript), https://www.ssrn.com/abstract=4997519.

5 *See* Tanina Rostain, *Robots Versus Lawyers: A User-Centered Approach*, 30 Geo. J. Legal Ethics 559, 571 (2017); David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. Pa. L. Rev. 1001, 1012 (2021).

6 Edwina L. Rissland, AI and Legal Reasoning 2 (1985).

7 *See* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 Ga. St. U. L. Rev. 1305, 1318 (2019).

8 Bojan Spaić & Miodrag Jovanović, *Artificial Reason and Artificial Intelligence: The Legal Reasoning Capabilities of GPT-4*, 72 Belgrade L. Rev. 383, 396 (2024); Harry Surden, *ChatGPT, AI Large Language Models, and Law*, 92 Fordham L. Rev. 1939, 1947 (2024).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

Somehow, training language models to predict the next word in a sequence has enabled them to competently perform a wide range of tasks that they were not explicitly trained to perform.[9] But it's hard to make heads or tails of what LLMs' legal reasoning capabilities are. Empirical research finds that LLMs can pass the bar exam,[10] outperform law students on law school exams,[11] and nearly match expert tax lawyers at answering tax questions.[12] But lawyers keep getting sanctioned—or at least publicly humiliated—for submitting ChatGPT-authored filings filled with bogus citations to nonexistent cases.[13] LLMs can write a limerick about language models' ability to write limericks but somehow can't recognize that the word *strawberry* has three *r*s in it.[14] Questions abound, but answers are scarce, largely because LLMs perform so differently than prior technology and because their capabilities are so far beyond what even their creators had anticipated.[15] What LLMs can do is astonishing. Given what LLMs can do, what they can't do is baffling.

This Article asks: Can the LLM breakthrough usher in a new era of automated self-help tools for low-income litigants?[16] To answer this, the Article examines three subsidiary questions. What are LLMs' legal reasoning capabilities? How do LLMs' strengths and weaknesses with legal reasoning shape the prospects for legal automation? Under what conditions would generative AI–powered legal automation expand or restrict access to justice?

LLMs' capabilities form what researchers call a "jagged frontier"—they excel at some tasks while failing at others of similar or lesser complexity.[17] Success in one context doesn't guarantee success in another, especially with reasoning tasks. Legal automation would seem to require an AI with a "general" legal reasoning skill: the ability to apply law to facts across different domains.[18] But LLMs don't think like lawyers. LLMs'

9 Jason Wei et al., Emergent Abilities of Large Language Models, 1 (Oct. 26, 2022) (unpublished manuscript), https://arxiv.org/abs/2206.07682.

10 Yonathan A. Arbel, *Judicial Economy in the Age of AI*, 25 U. Col. L. Rev. 549, 561 (2024).

11 Jonathan H. Choi & Daniel Schwarcz, AI Assistance in Legal Analysis: An Empirical Study 1 (Aug. 13, 2023) (unpublished manuscript), https://www.ssrn.com/abstract=4539836.

12 John J. Nay et al., *Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence*, 1 (Minn. Legal Stud. Rsch. Paper No. 23-15, Nw. Pub. L. Rsch. Paper No. 23-48, 2023), https://www.ssrn.com/abstract=4476325.

13 Frank Fagan, *A View of How Language Models Will Transform Law*, 92 Tenn. L. Rev. 1, 50 (2025); Sara Merken, *AI "Hallucinations" in Court Papers Spell Trouble for Lawyers*, Reuters (Feb. 18, 2025, 3:55 PM), https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/.

14 Katie Szilagyi, Regenerating Justice: ChatGPT and the Legal Minefield of Generative AI 12 (Nov. 14, 2024) (unpublished manuscript), https://www.ssrn.com/abstract=4976738.

15 Stephen Ornes, *The Unpredictable Abilities Emerging from Large AI Models*, Quanta Mag. (Mar. 16, 2023), https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/.

16 And can it do so without needing programmers to manually formalize the law into computer code.

17 Fabrizio Dell'Acqua et al., *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* 1 (Harv. Bus. Sch., Working Paper 24-013, 2023), https://www.ssrn.com/abstract=4573321.

18 *See* Jens Frankenreiter & Julian Nyarko, *Natural Language Processing in Legal Tech, in* Legal Tech and the Future of Civil Justice 72 (David Freeman Engstrom ed., 2023).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

proficiency with legal reasoning follows the same "jagged frontier." Depending on the task, an LLM might draft a winning argument or make nonsensical errors that no human would commit.[19] LLMs lack a "general" legal reasoning skill that would enable them to competently think through legal rules and fact patterns across different contexts.[20] LLMs generate text, one word at a time, by reproducing language patterns—not by reasoning in a structured way.[21] Given that LLMs are literally just predict-the-most-likely-next-word machines, their achievements are remarkable. But these accomplishments are also fragile. LLMs sometimes answer legal questions correctly even though their process of generating answers sidesteps actual legal analysis. But this power is limited to the model having been trained on many similar examples that provide an accurate pattern for the LLM to imitate.[22]

LLMs' lack of a "general" legal reasoning skill might suggest that the current generation of AI will not prove capable of automating legal reasoning work. This Article contends that generative AI's peculiarities should not lead to dismissing its potential for automating legal work altogether. Rather, our conception of legal automation should be reframed. Automation doesn't require machines to mimic human lawyers and possess the same skills. Mechanization has rarely involved machines that possess the same skills as humans, mimicking human action to produce identical products as human labor.[23] Automation typically involves machines performing narrow tasks more efficiently than humans. Because LLM capabilities form a jagged frontier, the frontier of potential automation is also jagged. Tasks within the frontier of LLM capabilities are ripe for automation, while tasks at the border or beyond are not. Because an LLM system's success in performing one legal reasoning task is not indicative of success in performing other tasks, automation requires extensive evaluation and iterative design for each specific task and domain.

Generative AI will either expand or restrict access to justice based upon where investments are made and how legal AI software is developed. Because legal automation requires

19 *See* Colin Doyle & Aaron D. Tucker, *If You Give an LLM a Legal Practice Guide, in* CSLAW '25: Proceedings of the 2025 Symposium on Computer Science and Law 194, 202, https://dl.acm.org/doi/10.1145/3709025.3712220.

20 *See id.* at 203; Sourav Banerjee et al., LLMs Will Always Hallucinate, and We Need to Live with This 8 (Sept. 9, 2024) (unpublished manuscript), https://arxiv.org/abs/2409.05746; Haiyan Zhao et al., *Explainability for Large Language Models: A Survey*, 15 ACM Transactions on Intelligent Sys. & Tech. 1, 19 (2024), https://dl.acm.org/doi/10.1145/3639372; Zikai Xie, Order Matters in Hallucination: Reasoning Order as Benchmark and Reflexive Prompting for Large-Language-Models 2 (Dec. 30, 2024) (unpublished manuscript), https://arxiv.org/abs/2408.05093.

21 Iman Mirzadeh et al., GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models 1 (Oct. 7, 2024) (unpublished manuscript), https://arxiv.org/abs/2410.05229.

22 Anton Korinek, *Generative AI for Economic Research: Use Cases and Implications for Economists*, 61 J. Econ. Lit. 1281, 1284 (2023).

23 Richard W. Hamming, The Art of Doing Science and Engineering: Learning to Learn 19 (2020).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

costly domain-specific and task-specific evaluation, the jagged frontier of legal automation will become increasingly jagged based upon where investments are made along the frontier. Access to justice could expand through targeted investment in legal AI software designed around the legal domains and issues of low-income litigants. But disproportionate investment in AI-driven tools that benefit high-income litigants at the expense of low-income litigants would exacerbate the access-to-justice crisis. The substantial investment needed to develop reliable legal AI software—and the financial incentive to keep this technology private—mean that those in positions of power have every incentive to develop bespoke AI tools that strengthen their advantage. Because it is so easy to build prototypes of generative AI software, there is a serious risk of proliferation of hastily built, unreliable legal aid software that misinforms indigent litigants. As software developed for the public good, legal aid self-help tools should be open source and transparent, not proprietary, both to ensure public accountability and to foster the development of industry standards for safety and reliability.

The Article makes three contributions. First, it diagnoses LLMs' legal reasoning capabilities. Second, it reframes the project of legal automation in light of those capabilities. Third, it assesses what conditions must exist for automation to expand access to justice. The Article proceeds as follows. Part I surveys the access-to-justice crisis, the turn to technological interventions as a potential solution, and the historical challenge of formalizing legal rules into code. Part II dissects LLMs' legal reasoning capabilities, examining the empirical literature on LLMs and legal reasoning and advancing the argument that LLMs' jagged legal-reasoning capabilities stem from their underlying architecture as pattern-matching machines. Part III considers the prospects of legal automation. Automation doesn't require LLMs to possess the same skills as human lawyers. The jagged frontier of LLMs' legal reasoning capabilities gives rise to a jagged frontier of possible automation. But LLMs' inherent unreliability will require extensive backend testing and iterative redesign. Part IV assesses automation's potential impact on access to justice, contending that investment choices will determine whether automation narrows or widens the justice gap. The benefits of the LLM breakthrough can be realized only through investment in software development. Open-source research focused on the legal issues of low-income litigants can expand access to justice, while proprietary tools risk deepening inequality.

I. ACCESS TO JUSTICE AND LEGAL TECHNOLOGY

A. Access-to-Justice Crisis

Millions of people in the United States face everyday legal problems—including evictions, debt collections, job terminations, denials of benefits or healthcare, and child custody matters—all without the benefit of professional legal assistance.[24] "In roughly

24 *See* Hagan, *supra* note 1, at 199.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

three-quarters of civil cases filed in state courts, at least one side lacks a lawyer."[25] The Legal Services Corporation's latest study on the justice gap finds that 74% of low-income Americans experienced civil legal problems in the past year but received adequate legal help for only 8% of these issues."[26] Many people don't seek legal help because they perceive the process as too costly, difficult, time-consuming, or unpleasant.[27] Low-income individuals often encounter barriers that prevent them from seeking legal assistance, including the inability to take time off work or afford childcare.[28] Others don't seek legal help because they don't recognize their problem as a legal problem or don't understand that legal assistance could help.[29] "To someone without money and without legal training, asserting legal rights is a mystery and an obstacle."[30]

People without lawyers fare poorly across family law, housing, and debt cases. Many Americans who cannot afford legal help routinely forfeit basic rights on important matters like debt collection, evictions, veterans' benefits, and child support. "They stay in bad marriages, pay exorbitant interest because of bad credit, [and] have their paychecks garnished for debts they don't owe."[31] Without legal assistance, "a person may lose their home to an eviction or foreclosure, suffer physical abuse without protection, go into debt or bankruptcy, lose custody of their children, or be denied medical care."[32]

The terms *justice gap* and *access-to-justice crisis* refer to the stark disparity between available legal services in the United States and the needs of the population, particularly poor communities.[33] The fundamental problem is simple: There aren't enough affordable legal services to meet public needs. The straightforward solution would be to provide more legal services through increased funding for legal aid organizations.[34] But this approach has not been politically viable. Courts and legislatures have not recognized a right to "civil Gideon"—a guarantee of legal representation in civil matters analogous to the right

25 David Freeman Engstrom, *Introduction: Civil Justice at the Crossroads, in* Legal Tech and the Future of Civil Justice 2 (David Freeman Engstrom ed., 2023).

26 Legal Servs. Corp., *supra* note 3, at 5–8.

27 *Id.* at 8; Katherine L. W. Norton, *The Middle Ground: A Meaningful Balance Between the Benefits and Limitations of Artificial Intelligence to Assist with the Justice Gap*, 75 U. Mia. L. Rev. 190, 195 (2020).

28 Legal Servs. Corp., *supra* note 3, at 5.

29 Hagan, *supra* note 1, at 200.

30 Martha Minow, *Access to Justice*, 2 Am. J.L. & Equal. 293, 296 (2022).

31 Rowan Moore Gerety, *Lawyer Up? Increasingly, Americans Won't, or Can't.*, N.Y. Times (Apr. 17, 2025), https://www.nytimes.com/2025/04/17/magazine/lawyers-civil-court.html.

32 Hagan, *supra* note 1, at 199.

33 *See* Legal Servs. Corp., *supra* note 3, at 5–8; Norton, *supra* note 27, at 233.

34 *See* Minow, *supra* note 30, at 295; Colleen V. Chien & Miriam Kim, *Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap*, 57 Loy. L. Rev. 903, 908 (2025) ("A major channel of legal assistance to low-income consumers is through state and philanthropically funded legal aid organizations, which represent individuals in consequential matters such as eviction defense, deportation proceedings, and criminal defense.").

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

defendants have in criminal proceedings.[35] Although Congress, state legislatures, and philanthropic organizations do fund legal aid organizations, the funding is inadequate for these organizations to meet the needs of the communities they serve.[36] Many legal aid organizations turn away half the requests they receive and fully resolve only half the cases they accept.[37]

Given the political reality that legal aid will be perpetually underfunded, courts, regulatory agencies, and public interest organizations have increasingly turned to technological solutions to bridge the justice gap.[38] These technologies aim to increase access to the legal system without the need for lawyers in two main ways: facilitating the litigation process and providing self-help tools.

Technology facilitates litigation by removing practical barriers to people pursuing litigation at a courthouse, such as the inability to take time off work, childcare obligations, and transportation challenges.[39] The most prominent example of this technology today is online dispute resolution, or ODR, which allows litigants to resolve their disputes through video-conferencing platforms rather than by physically appearing in court.[40] Self-help tools aim to equip individuals without legal representation with the knowledge and tools necessary to navigate the legal system on their own.[41] Broadly speaking, these resources provide potential litigants with information to help them understand whether their conflict constitutes a legal issue, recognize their rights, identify legally relevant facts, navigate court procedures, and evaluate their available options.[42] Common implementations include reports, pamphlets, and websites written in plain language and designed to clarify

35 Norton, *supra* note 27, at 217–18.

36 Chien & Kim, *supra* note 34, at 908; Legal Servs. Corp., *supra* note 3, at 9.

37 Legal Servs. Corp., *supra* note 3, at 9.

38 Rostain, *supra* note 5, at 571.

39 J. J. Prescott, *Using ODR Platforms to Level the Playing Field: Improving Pro Se Litigation Through ODR Design, in* Legal Tech and the Future of Civil Justice 287 (David Freeman Engstrom ed., 2023).

40 Eboigbe, *supra* note 4, at 25. These platforms can handle a wide range of disputes, from e-commerce issues to family law matters. *Id.* These virtual courtrooms, which ballooned during COVID, when in-person appearances became impossible, have endured since the pandemic subsided. Engstrom, *supra* note 25, at 1001; Victor D. Quintanilla et al., *Digital Inequalities and Access to Justice: Dialing into Zoom Court Unrepresented, in* Legal Tech and the Future of Civil Justice 225 (David Freeman Engstrom ed., 2023). This model provides significant convenience and flexibility, alleviating the burdens associated with traditional litigation. But online proceedings are accessible only to those with adequate technology, reliable internet, and digital literacy. *See generally* Rebecca L. Sandefur, Legal Tech for Non-Lawyers: Report of the Survey of US Legal Technologies (2019) (finding that legal technology tools fail to improve access to justice for low-income communities because of barriers to using the technology).

41 Norton, *supra* note 27, at 239; Samuel Dahan & David Liang, *The Case for AI-Powered Legal Aid*, 46 Queen's L.J. 415, 427 (2021).

42 *See* Rostain, *supra* note 5, at 12; Engstrom & Gelbach, *supra* note 5, at 11.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

legal issues and processes for laypersons.[43] "Read only" self-help tools have inherent limitations. Users must recognize that their issue is legal in nature, identify the correct self-help resources to use, and find the answer to their legal problems within those materials. The task of navigating dense legal information can be daunting, overwhelming users who are unfamiliar with legal concepts or who lack the technical proficiency to navigate court and legal aid websites.[44] The ideal solution is automation: digital systems that don't make users comb through documents or parse dense text, but instead take information about a user's situation and provide tailored legal information or guidance.[45]

B. Automated Legal Assistance

The challenge with developing automated systems for legal reasoning is substantial: It requires formalizing law into code. For decades, computational law has focused on developing logic-based systems that simulate legal analysis through a series of logical steps for a computer to execute.[46] This requires not merely translating legal rules into code but also encoding the way that the law should apply to particular facts—in other words, legal reasoning itself. The resulting systems comprise complex decision trees with intricate if-then logic designed to navigate from a set of facts to a legal conclusion. Creating these systems requires enormous effort from legal and technical experts working together. To use this software, litigants typically answer multiple-choice questions that function as decision trees or flowcharts leading to specific information or advice.[47] But even well-designed

43 As one example, consider the Massachusetts Legal Resource Finder, which provides targeted contact information and self-help materials to those seeking legal assistance. Norton, *supra* note 27, at 195.

44 *Id.* at 210, 227.

45 Automation of legal reasoning should be understood to include not just the provision of legal services but also the provision of legal information. Computer systems that provide relevant legal information still require legal reasoning, even if these systems don't draw legal conclusions by directly applying the law to the facts of user's case. These systems need to process the raw information of a user's case and identify the facts that are legally relevant. Much of the challenge of access to justice stems from people not understanding their problems as legal problems. An identification of relevant legal information requires conceiving of the facts of a given case within a legal framework.

46 Andrew Coan & Harry Surden, *Artificial Intelligence and Constitutional Interpretation* 9 (Ariz. Legal Stud. Discussion Paper No. 24-30; U. Col. L. Legal Stud. Rsch. Paper No. 24-39, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5018779. Computational models of legal reasoning have been developed both to better understand legal reasoning by simulating it in a computer program and to study how computers can aid legal practice by automating legal reasoning tasks. Henry Prakken, *Legal Reasoning: Computational Models, in* International Encyclopedia of the Social & Behavioral Sciences 784, 784 (James D. Wright ed., 2015).

47 While most of these systems have been developed by courts and legal aid organizations, the private market for individual legal services has also seen a surge in activity, with companies developing automated web-based processes to produce customized basic legal documents such as wills, divorce agreements, contracts, and incorporation papers without the direct involvement of a lawyer. Dana Remus & Frank S. Levy, Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law 3 (Dec. 11, 2015) (unpublished manuscript), https://www.ssrn.com/abstract=2701092; Eboigbe, *supra* note 4, at 25. These tools are particularly useful in straightforward areas of law, such as family and landlord–tenant law, and can also assist users in navigating complex regulatory regimes, such as disaster assistance regulation. Rostain, *supra* note 5, at 571.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

systems can frustrate users, who may struggle to understand the questions or match their circumstances to the multiple-choice options that the system presents. This is a particularly serious problem, as unrepresented litigants struggle to frame their disputes in legal terms and are seeking help from these tools to do so.[48]

Legal reasoning is messy, humanistic, linguistic, nondeterministic, contradictory, ambiguous, scruffy, and just plain hard to pin down—in other words, precisely the kind of work that computers can't do.[49] Because logic-based systems rely on hand-coded expert rules, they are "brittle" and struggle to handle exceptions, non-standard scenarios, conflicts, discretion, and nuances.[50] These systems have been critiqued for struggling with the complexity and ambiguity of legal language, often watering down legal concepts.[51] Constrained by their predefined rules, these systems lack the flexibility and adaptability inherent in legal reasoning.[52] Analogical arguments require the ability to draw appropriate analogies between different fact patterns, but identifying relevant similarities between distinct sets of facts depends on the principle for which the initial case stands, which involves evaluative judgments.[53] Determining specific legal principles from case law is also difficult because the text of a decision may not clearly present the nature of the legal reasoning involved.[54] And legal opinions often rely on unstated background knowledge, including moral and political theories, intuition, biases, and common-sense reasoning.[55]

Getting computers to perform legal reasoning tasks has been an insurmountable challenge because of the demands of formalization: translating law and legal reasoning from

48 People skeptical of these tools argue that AI could interfere with the legal profession, potentially violating professional conduct rules or constituting unauthorized practice of law. Norton, *supra* note 27, at 195; Chien & Kim, *supra* note 34, at 920.

49 Eboigbe, *supra* note 4, at 7.

50 *Id.*

51 *Id.*

52 Legal domains with open-textured rules like multifactor tests are not susceptible to this kind of modeling. Other legal rules cannot be easily translated into executable code because of their ambiguous and evolving nature. Rissland, *supra* note 6, at 1. Rules derived from precedents are often specific to cases, making it difficult to apply them to new cases. Prakken, *supra* note 46, at 789. Because rules from case law are rarely rigidly defined and depend on continual interpretation and reinterpretation, it's difficult to formalize the necessary and sufficient conditions for applying legal predicates. Rissland, *supra* note 6, at 2.

53 *See* Cass R. Sunstein, *Of Artificial Intelligence and Legal Reasoning*, 8 U. CHI. L. SCH. ROUNDTABLE 29, 32–33 (2001) (critiquing legal AI systems that made analogical arguments based on factual similarities without considering a principle that would make those facts legally significant). Analogical reasoning requires taking a large body of discrete cases and indexing, generalizing, and applying information from those cases. RISSLAND, *supra* note 6, at 1.

54 Dean Alderucci, *The Automation of Legal Reasoning: Customized AI Techniques for the Patent Field*, 58 DUQ. L. REV. 50, 68 (2020).

55 *Id.*

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

human texts and practices into computer code that a machine can execute.[56] These systems require a tremendous amount of human labor to set up. A domain expert must manually—and accurately—program the many interlocking rules that the software must evaluate to arrive at a correct conclusion. When rules conflict, methods need to be programmed to allow the software to resolve those conflicts, which only increases the complexity of these systems and the labor required to manage them.[57] Inherent in the design of rule-based systems is an assumption that the law is determinate and calculable, which leaves no room for legal issues that are indeterminate with potentially valid but contradictory legal arguments.[58] Unlike mathematical proofs focused on correctness, legal arguments aim to persuade, and the correctness of a legal argument often cannot be strictly defined.[59] In practice, these logic-based systems have gained the most traction within domains of law that are heavily logical and have precise rules and clear applications.[60]

Since the 1980s, scholars in the AI and law field have speculated about overcoming this formalization hurdle through an AI possessing a legal reasoning ability.[61] Like a human lawyer, this AI system could take a legal text, develop a conceptual understanding, and apply that understanding to a set of facts. This AI would eliminate the need for experts to meticulously program logical structures for every legal domain and fact pattern. Instead, you would simply give the AI the relevant law, and it would process the facts of a dispute to reach appropriate legal conclusions.

But what does *legal reasoning* even mean? Scholars have debated this for over a century, with no consensus in sight.[62] Some contend that legal reasoning is distinctive and different from logic, scientific reasoning, or ordinary decision-making.[63] Others see it as

56 *See* Dahan & Liang, *supra* note 41, at 8; Sarah Lawsky, *Coding the Code: Catala and Computationally Accessible Tax Law*, 75 SMU L. Rev. 535, 536 (2022) ("Even today, automated formalizations such as natural language processing have yet to succeed in translating tax law into computer code or other formalizations.").

57 Prakken, *supra* note 46, at 789.

58 Lance B. Eliot, AI and Legal Argumentation: Aligning the Autonomous Levels of AI Legal Reasoning 4 (Sept. 11, 2020) (unpublished manuscript), https://arxiv.org/abs/2009.11180. Although a rule-based approach may be able to address "easy" cases, it's an insufficient basis for developing legal arguments for hard cases, as these kinds of legal arguments are not purely logical but depend upon rhetoric designed to persuade others to accept a particular interpretation of the law. *Id.*

59 *Id.*

60 A prototypical example is tax-preparation software such as TurboTax, where tax laws are translated into formal rules that a computer can process. Surden, *supra* note 7, at 1318.

61 Rissland, *supra* note 6, at 1.

62 Phoebe C. Ellsworth, *Legal Reasoning, in* The Cambridge Handbook of Thinking and Reasoning 685 (Keith J. Holyoak & Robert G. Morrison Jr. eds., 2005); Larry Alexander & Emily Sherwin, *Demystifying Legal Reasoning: Part II* 1 (Cornell L. Sch. Rsch. Paper No. 06-002, 2006).

63 Ellsworth, *supra* note 62, at 685.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

ordinary reasoning applied to legal problems.[64] Skeptics believe that ideologies, politics, biases, and policy preferences matter more for legal decision-making than the reasoning explicitly laid out on the pages of legal documents.[65]

Law remains a pragmatic field and profession. The lack of a precise definition of *legal reasoning* hasn't stopped generations of lawyers and judges from doing whatever it is that legal reasoning is, even if no one agrees on what that is. Machines might follow human lawyers' footsteps. Debates over legal reasoning tend to concern its outer bounds or its ontological status as a unique form of reasoning.[66] Although a precise definition of *legal reasoning* is debated, everyone understands legal reasoning as a syncretic form of reasoning that balances logical, contextual, moral, social, practical, and political concerns.[67] As a day-to-day matter—particularly in this Article's context of routine disputes over housing, family law, and small claims issues—legal reasoning can be understood as applying law to facts. A legal-reasoning skill would reflect an actor's capacity to take a given set of facts and understand their significance given a set of legal rules.

The skill of applying law to facts is still not as simple as it might appear. It involves special skills that we don't typically associate with computers, including understanding complex ideas expressed in natural language, balancing different methods of reasoning and argumentation, and writing persuasive arguments in prose form.[68] In early AI and law research, Edwina L. Rissland identified six capabilities any AI legal reasoning system would need:

> (1) the ability to reason with cases and examples, particularly through analogy; (2) the ability to handle ill-defined, open-textured predicates; (3) the ability to handle exceptions; (4) the ability to handle fundamental conflicts between rules; (5) the ability to argue and justify in a case-based manner; and (6) the ability to handle change and nonmonotonicity.[69]

64 Alexander & Sherwin, *supra* note 62, at 1 (clarifying that "ordinary reasoning" is understood to include moral reasoning, empirical reasoning, and deduction from authoritative rules).

65 *See* Frederick Schauer, Thinking Like a Lawyer: A New Introduction to Legal Reasoning 150 (2009) (summarizing this viewpoint). According to this view, judges often make decisions based on nonlegal factors, such as personal characteristics, fairness judgments, or policy considerations, rather than strictly adhering to legal doctrine. *Id.* The persistence of legal doctrine is a result of judges needing a "doctrinal hook" to justify decisions made on these broader bases. *Id.* at 161.

66 *See* Ellsworth, *supra* note 62, at 685; Alexander & Sherwin, *supra* note 62, at 1.

67 Legal reasoning is only pseudo-formalized, with a large body of rules that can be ambiguous, contradictory, and incomplete. Ellsworth, *supra* note 62, at 671. Lawyers and judges weigh competing principles, consider policy implications, and draw upon practical wisdom. Ray Worthy Campbell, *Artificial Intelligence in the Courtroom: The Delivery of Justice in the Age of Machine Learning*, 18 Colo. Tech. L.J. 323, 340 (2023).

68 *See* Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 Yale L.J. 1957, 1961 (1990).

69 Edwina L. Rissland, *Artificial Intelligence and Legal Reasoning*, AI Mag., fall 1988, at 45, 48.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

Until recently, this vision of computer software possessing legal reasoning skills remained firmly in the realm of science fiction. Computers simply couldn't work with natural language or engage in open-ended thinking.[70] The breakthrough of LLMs marks the point in time at which the prospect of computers performing legal reasoning transformed from a niche research topic to a mainstream concern for the legal industry.[71] LLMs exhibit a remarkable felicity with written language. As Jens Frankenreiter and Julian Nyarko noted, "The legal system trades in words, and [natural language processing with LLMs] promises to automate an activity that lies at the heart of many tasks performed by lawyers: the extraction and processing of information from unstructured text."[72] Generative AI offers the tantalizing possibility of automating legal reasoning without formalization.

LLMs upend our expectations about what computers can do, how they should act, and what kind of mistakes they're prone to make. Contemporary LLMs demonstrate remarkable feats of legal reasoning: they can pass the bar exam, outperform law students on examinations, and even hold their own against tax lawyers when answering specialized tax questions.[73] But these same models are prone to serious—sometimes catastrophic—errors, such as hallucinating fictitious case citations and making obvious, nonsensical reasoning mistakes.[74] This raises serious questions about whether generative AI can truly overcome the formalization hurdle that has stymied the project of legal automation.

C. *Technology and Inequality*

If generative AI proves capable of automating legal reasoning work, will this usher in a new era of expanded access to justice, or will existing inequities get worse?[75] The law and technology literature presents two competing visions for how AI will impact access to justice: an optimistic view and a pessimistic view. Optimists believe in AI's potential to democratize access to legal knowledge and advice. Pessimists fear that AI will compound existing advantages and thereby deepen inequality.

The optimistic view emphasizes technology's potential to lower litigation costs and connect people with services and information. By making legal services more accessible and affordable, technology could improve access to justice, either as a targeted intervention

70 Rissland, *supra* note 68, at 1967.

71 Jinqi Lai et al., *Large Language Models in Law: A Survey*, 5 AI Open, 181, 181 (2024).

72 *See* Frankenreiter & Nyarko, *supra* note 18, at 71.

73 Spaić & Jovanović, *supra* note 8, at 399; Nay, *supra* note 12, at 11.

74 Matthew Dahl et al., Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models 1 (Jan. 2, 2024) (unpublished manuscript), https://arxiv.org/abs/2401.01301.

75 The conclusions of parts II and III can dismiss an important third possibility: that generative AI is nothing more than a Silicon Valley parlor trick masquerading as a world-changing technological disruption, a dud with little impact on access to justice.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

or as a downstream product of broader technological developments.[76] As Chief Justice John Roberts noted in his *2023 Year-End Report*,

> [f]or those who cannot afford a lawyer, AI can help. It drives new, highly accessible tools that provide answers to basic questions, including where to find templates and court forms, how to fill them out, and where to bring them for presentation to the judge—all without leaving home. These tools have the welcome potential to smooth out any mismatch between available resources and urgent needs in our court system.[77]

The optimists predict that AI will improve access to justice in two ways: automation and augmentation.[78] With automated systems, many legal services would be delegated to software systems with minimal human involvement or oversight. Lawyers would step in to handle tasks beyond the automated systems' capabilities. In contexts in which automation is infeasible, legal technology might still be able to cut the costs of legal services through "intelligence augmentation," by which technology enhances lawyers' capabilities.[79] Either pathway—automation or augmentation—would increase access to justice, the first by empowering clients for self-help and the second by reducing the costs of legal services.[80]

Few optimists see technology as a panacea for access-to-justice problems capable of achieving the goal of providing adequate legal services to all civil litigants.[81] But they see technology as a tool that can help. And they worry that a myopic focus on the harms of AI will prevent society from capturing its benefits, particularly for access to justice.[82]

The pessimists predict that AI will exacerbate current inequalities. They acknowledge that legal technology might streamline processes and make litigation more efficient, thereby increasing access to justice by removing or lowering barriers to litigation for

76 *See* Engstrom & Gelbach, *supra* note 5, at 1011; Remus & Levy, *supra* note 47, at 3. AI may not only reduce the cost of legal services but may also address social and psychological barriers that prevent people from seeking legal recourse. Arbel, *supra* note 10, at 566.

77 2023 Year-End Report on the Federal Judiciary 5, U.S. Sup. Ct. (2023), https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf.

78 An analogous example is AI for driving cars. An automated self-driving car system would operate independent of a human driver, perhaps calling in for human assistance for special circumstances. An AI that augments a human driver might do things like provide alerts, temporarily redirect the vehicle to stay in a lane, or slam on the brakes if it detects an imminent front-end collision.

79 Taylor Poppe, *The Future Is Complicated: AI, Apps & Access to Justice*, 72 Okla. L. Rev. 185, 189 (2019).

80 *Id.*

81 Norton, *supra* note 27, at 232.

82 Orly Lobel, The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future 5 (2022).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

low-income communities.[83] But they believe that technology's costs will outweigh its benefits. The primary effect of technology will be to increase existing asymmetries in litigation. If technology provides the "haves" in litigation an even greater advantage over the "have nots," technology's net effect will be to widen the gap in access to justice.[84] Although AI tools are becoming more widely available, wealthy clients and firms will have access to more sophisticated systems and proprietary data sources.[85] By investing in private AI tools that glean insights from troves of privately held data, large firms will be able to use AI to further capitalize on their advantage,[86] concentrating power rather than democratizing access to justice.[87]

II. LANGUAGE MODELS' LEGAL REASONING CAPABILITIES

Can LLMs' ability to manipulate language and generate text that appears human extend to the complex, open-ended task of legal reasoning? Understanding LLMs' legal reasoning capabilities requires a conceptual understanding of how LLMs are constructed and how they operate. Part II.A is a primer on how generative AI works and how language models are incorporated into legal AI software. Part II.B is an analysis of LLMs' legal reasoning capabilities and an explanation of their shortcomings. The core challenge is that LLMs do not perform structured reasoning.[88] They can excel at legal reasoning tasks that are well-represented within their training data because their training data provides them with a language pattern to follow and imitate.[89] The bottleneck on what LLMs can accomplish is the underlying logic that LLMs follow: a semantic logic that detects and replicates language patterns.

83 Engstrom & Gelbach, *supra* note 5, at 1004. Others have voiced concerns that the digitization and commoditization of legal guidance could result in the loss of essential values and functions that the American legal profession upholds, which are crucial to democracy and societal functions. *See* Raymond H. Brescia, *What's a Lawyer For? Artificial Intelligence and Third-Wave Lawyering*, 51 Fla. St. U. L. Rev. 543, 545 (2024).

84 Engstrom & Gelbach, *supra* note 5, at 1004.

85 Harry Surden, *Machine Learning and Law: An Overview, in* Research Handbook on Big Data Law 10 (Roland Vogl ed., 2021).

86 David Freeman Engstrom & Nora Freeman Engstrom, *Legal Tech and the Litigation Playing Field, in* Legal Tech and the Future of Civil Justice 135 (David Freeman Engstrom ed., 2023) (finding that "repeat players are already using their tech savvy and their stranglehold on confidential claims data to drive case settlements downward").

87 Nick Noonan, Creative Mutation: A Prescriptive Approach to the Use of ChatGPT and Large Language Models in Lawyering 32 (Mar. 13, 2023) (unpublished manuscript), https://www.ssrn.com/abstract=4406907.

88 Subbarao Kambhampati et al., *Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks* 1 (Int'l Conf. on Machine Learning, 2024), https://arxiv.org/pdf/2402.01817; Kyle Mahowald et al., Dissociating Language and Thought in Large Language Models 11 (Mar. 23, 2024) (unpublished manuscript), https://arxiv.org/abs/2301.06627; Mirzadeh et al., *supra* note 21, at 2.

89 Korinek, *supra* note 22, at 4.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

A. How Generative AI Works

LLMs belong to the field of machine learning, which is a "family of AI techniques" that analyzes large datasets, learns patterns, and assembles knowledge from that data.[90] The "learning" in machine learning is metaphorical. It refers to the system's ability to improve its performance over time by independently adjusting a statistical model to fit a dataset, not to any form of understanding or cognition.[91] Machine learning has largely overtaken logic-based systems as a more efficient, powerful way to build software to perform tasks associated with human thinking. With machine learning systems, rules are not predefined by a programmer. Instead, the software independently develops its own rules.[92]

How is it that LLMs can produce text that feels like it was written by a person? Language models' understanding of language is not built through explicitly programmed rules about language but through machine learning methods that glean statistical patterns from trillions of examples of how humans use language.[93] Language models' impressive capabilities can be attributed to three features of modern LLMs: (1) the architecture of the machine learning methods that process text data, (2) the vast amount of training data,[94] and (3) a final stage in the training process that calibrates the LLM to respond to human instruction.[95]

When generating text, LLMs operate through a surprisingly simple mechanism: predicting one small chunk of text at a time.[96] These small chunks of text, called tokens, are the basic units the model processes. Each token represents a word, a part of word, individual characters, or even punctuation marks.[97] Take a simple sentence: "The law is

90 Surden, *supra* note 7, at 1311.

91 Colin Doyle, *How Algorithms Expose the Law* 72 (Harv. Pub. L. Working Paper No. 21-45, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3894617.

92 Coan & Surden, *supra* note 46, at 9. Machine learning offers many advantages over logic-based programming for artificial intelligence. Logic-based systems follow explicitly defined rules to turn input variables into output variables. Logic-based hard-coded systems can be understood as a top-down approach to artificial intelligence, and machine learning systems as a bottom-up approach to artificial intelligence. John Armour et al., *Augmented Lawyering*, 2022 U. Ill. L. Rev. 71, 85 (2022). Compared to logic-based systems, machine learning models can achieve remarkable results with much less human labor involved. Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 Duke L. & Tech. Rev. 18, 26 (2017). Instead of requiring a programmer to code by hand an endless number of if-then structures, with machine learning, the computer software learns the underlying relationships and builds a structure on its own.

93 *Oversight of A.I.: Principles for Regulation*, U.S. Senate Comm. on the Judiciary (2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation.

94 Ethan Mollick, *Thinking Like an AI*, One Useful Thing (Oct. 20, 2024), https://www.oneusefulthing.org/p/thinking-like-an-ai.

95 Korinek, *supra* note 22, at 4.

96 Miguel Grinberg, *How LLMs Work, Explained Without Math*, miguelgrinberg.com (May 5, 2024), https://blog.miguelgrinberg.com/post/how-llms-work-explained-without-math.

97 *Id.*

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

complex." A large language model might break those words into tokens: "The," "law," "is," and "complex." To enable a computer to process the tokens, they are converted into unique numbers representing the text.[98] Where we see "The law is complex," an LLM sees something like "345, 678, 123, 890." At each step of generating new text, the model examines the tokens that have come before—both the text that the user has written as a prompt and any text that the LLM has already generated—and calculates what token is statistically most likely to come next.[99] Given the text "The law is," a model will assign probabilities to possible next tokens:

"complex": 70% likely

"clear": 20% likely

"unjust": 10% likely

The LLM then selects the most probable token (or, sometimes, randomly samples from these possibilities to add variety).[100] To build up a passage of text, this process repeats, generating text one token at a time, with each new token becoming part of the context for the next prediction.

If all LLMs do is predict the next token in a sequence, how do these predictions produce meaningful text? LLMs are built on a neural network architecture, which is loosely inspired by the structure and function of the human brain. In a neural network, layers of artificial neurons process information sequentially.[101] Each neuron receives input from other neurons, applies mathematical transformations, and passes results forward.[102]

Think of neural networks as a giant game of Plinko, the carnival game featured on the TV show *The Price Is Right* where you drop a disc from the top of a board and watch it bounce off pegs until it lands in a prize bin at the bottom. The game is random. While you choose where to drop the disk, its path downward is largely left to chance. The Plinko board is like a neural network's architecture: The disk is input data, the pegs are the nodes or neurons that the data passes through, and the bins are the model's potential outputs. Drop a disk onto a Plinko board and the disk will pinball around to land on a random result. That's exactly what happens at the start of training an LLM. When a neural network begins the training process its parameters are random, so feeding it data produces random results. An untrained language model can generate only gibberish. Ask an

98 *Id.*

99 *Id.*

100 Grinberg, *supra* note 96; Surden, *supra* note 8, at 1951.

101 Sylvain Christin et al., *Applications for Deep Learning in Ecology*, 10 Methods Ecology & Evolution 1632, 1634 (2019).

102 *Id.*

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

untrained language model to complete the sentence "The cow says . . ." and it is as likely to generate the word "inhibition" or "photosynthesis" as it is to generate the word "moo."

Now imagine that you could tweak the position of each peg on the Plinko board after every drop, nudging the pegs bit by bit to guide differently shaped disks toward the right bin. That's essentially how neural networks learn. During training, the model repeatedly practices predicting what word comes next in a sequence and compares its prediction against a real-world example.[103] When the network makes a wrong prediction, the model ever so slightly adjusts those Plinko pegs.[104] After doing this literally billions of times, the disks start to consistently land where they should. The network has gradually transformed from producing gibberish to producing coherent results. Through a mind-bogglingly lengthy process, the model incrementally develops a statistical understanding of how words relate to each other.[105] This understanding is not captured as explicit rules or instructions to follow but is implicit, embedded within the patterns of the neural network itself.

To fully capture the relationships between words, LLMs need to be trained on massive datasets drawn from diverse text sources; books, articles, websites—essentially the entire internet.[106] Here, the Plinko-board metaphor falls apart.[107] While a carnival Plinko board might have a few dozen pegs, an LLM's metaphorical board has billions of them arranged in thousands of layers of interconnected Plinko boards. Instead of a disk falling in two dimensions, the model's "disk" traverses a multidimensional space of billions of nodes. This massive scale of data, internal parameters, and training time is what enables these models to capture the statistical relationships between words across the English language.[108]

Another reason the text that modern language models produce feels more human and less robotic than previous generations of language models is that LLMs are more attuned to contextual clues. The "transformer" architecture that powers modern LLMs lets the model focus on relevant parts of the input text.[109] Transformers employ a mechanism called "self-attention," which allows the model to weigh the importance of different parts

103 Surden, *supra* note 8, at 1959.

104 The neural network's parameters.

105 Péter Homoki & Zsolt Ződi, *Large Language Models and Their Possible Uses in Law*, 64 HUNGARIAN J. LEGAL STUD. 435, 436–38 (2024); Yonathan A. Arbel & David Hoffman, Generative Interpretation 27 (Aug. 14, 2023) (unpublished manuscript), https://arxiv.org/abs/2308.06907.

106 Zhao et al., *supra* note 20, at 11. Although details about the training datasets for LLMs are kept secret, it is estimated that current LLMs have been trained on trillions of words. Korinek, *supra* note 22, at 4.

107 This is why we can't have actual mechanical language models.

108 Arbel & Hoffman, *supra* note 105, at 30; Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 610, 610–12 (Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency), https://dl.acm.org/doi/10.1145/3442188.3445922.

109 Surden, *supra* note 8, at 1962; Arbel & Hoffman, *supra* note 105, at 27.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

of the input text when making predictions. Unlike earlier systems that analyzed words in strict sequence, transformer models can consider all words in a passage simultaneously, weighing their relationships to each other.[110]

The final step in building an LLM is fine-tuning.[111] This process recalibrates the model so it will be responsive to human instruction, behave more like an assistant, and avoid generating harmful or inappropriate content.[112] Different methods are used at this stage, but they are all highly labor intensive. Human beings compare, rate, and flag responses that the LLM generates, and the model is adjusted according to each piece of feedback.[113] This calibration step is necessary to produce language models that follow instructions, stay on topic, and produce complex texts.[114]

110 Surden, *supra* note 8, at 23; *Transformers Laid Out*, Pramod's Blog (Jan. 3, 2025), https://goyalpramod.github.io/blogs/Transformers_laid_out/. Consider how a human reads the sentence "The judge who wrote the landmark opinion retired last year." Even though the word "opinion" can have multiple meanings, a human reader would infer that "opinion" in this sentence must refer to a formal legal decision. Traditional neural networks struggled with these kinds of connections, but the attention mechanism allows language models to link relevant words together. The model accomplishes this by calculating scores for relationships between each word and every other word in the input text based on their surrounding context. Arbel & Hoffman, *supra* note 105, at 27. These scores determine how much each word in a sequence should influence the model's understanding of other words in the sequence. In the sample sentence, the words "judge," "wrote," and "landmark" would inform the meaning of the word "opinion." These importance scores shift with each new input, allowing the model to capture slight variations in meaning.

111 Orly Mazur & Adam Thimmesch, Beyond ChatGPT: Transforming Government with Augmented LLMs 13 (June 27, 2024) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4846472; Cameron R. Wolfe, *A Brief Summary of Language Model Finetuning*, Stack Overflow (Oct. 31, 2024), https://stackoverflow.blog/2024/10/31/a-brief-summary-of-language-model-finetuning/.

112 Mazur & Thimmesch, *supra* note 111, at 13; Hyung Won Chung et al., Scaling Instruction-Finetuned Language Models 11 (Dec. 6, 2022) (unpublished manuscript), https://arxiv.org/abs/2210.11416.

113 April G. Dawson, *Algorithmic Adjudication and Constitutional AI—The Promise of a Better AI Decision Making Future?*, 27 SMU Sci. & Tech. L. Rev. 11, 30–31 (2024).

114 Most of us never interact with a base model before this fine-tuning phase, but understanding the base model can help us understand the underlying limitations that persist after fine-tuning. If you ask a fine-tuned LLM like ChatGPT "What is your name and what do you do?" it will respond with something like "I'm ChatGPT, your AI assistant. My role is to help you with tasks, answer questions, provide guidance, and collaborate on projects across a variety of topics. Whether you need help brainstorming ideas, solving problems, managing workflows, or just having a conversation, I'm here for you! ☺" If you ask an LLM that has not been fine-tuned "What is your name and what do you do?" it will respond with something like "My name is David and I am a 3rd year student at the University of Toronto. I am a student in the Faculty of Arts and Science, and I am majoring in Political Science and minoring in History." That's because this base model is *only* a next-word prediction machine. It takes the input text "What is your name and what do you do?" and then, token by token, fills out the text that is statistically most likely to follow based on its training data. The LLM isn't lying to you about being a University of Toronto student named David. It's just doing its job of filling out the next part of a text sequence with the text that's most likely to come after the text you've given it. Fine-tuning is a process of adjusting the model so it doesn't just autocomplete the text it's been given but treats aspects of that text differently, such as direct questions and instructions on what the LLM should do. *See* Wolfe, *supra* note 111.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

1. ***Legal AI Software***—Legal AI software works like a car built around an LLM engine. Legal AI software developers do not build their own language models. Instead, they take existing LLMs—such as OpenAI's ChatGPT or Google's Gemini—and build software components around them. Car manufacturers build structures around a combustion engine to enhance performance for certain tasks: spoilers to keep NASCAR race cars hugging the road and elevated chassis to lift pickup trucks up from the road. But a car remains bound by its engine's capabilities. Even the fastest Ferrari cannot break the sound barrier. Similarly, software developers build structures around an LLM to align with different goals. These structures can enhance performance of certain legal tasks but cannot transcend the underlying LLM's performance.

A handful of methods for improving LLM-powered software have emerged: fine-tuning, reinforcement learning with human feedback, retrieval-augmented generation, function calling and tool use, and prompt engineering.

Fine-tuning transforms general-purpose language models into specialized tools.[115] The process takes an existing LLM and trains it on a specialized dataset for a particular domain, updating a portion of the model's parameters to follow the language patterns of this specialized dataset when generating text.[116] For legal AI software, this means fine-tuning an LLM on a dataset of legal texts. Compared to a general-purpose language model that has been trained to follow the patterns of all of the text on the internet, the responses of this fine-tuned model will more closely follow the language patterns of legal writing. Fine-tuning allows developers to piggyback on the power of existing models rather than build legal AI systems from scratch—an enormously expensive and time-consuming task.[117]

Reinforcement learning with human feedback—referred to as "RLHF"—means that human experts manually guide a model toward better performance by providing feedback on the LLM's responses.[118] Human experts evaluate an LLM's outputs, often selecting better responses from alternatives. These evaluations create a dataset that trains a "reward model" to capture human preferences.[119] If you give this model a piece of text, it will rate how much that text conforms to the preference of human experts. Going forward, the LLM and reward model work together to generate text.[120] When generating text, the LLM produces multiple potential responses, the reward model rates them, and the LLM

115 Mazur & Thimmesch, *supra* note 111, at 13.

116 Young-Suk Lee et al., AMR Parsing with Instruction Fine-tuned Pre-trained Language Models 6 (Apr. 24, 2023) (unpublished manuscript), https://arxiv.org/abs/2304.12272.

117 Mazur & Thimmesch, *supra* note 111, at 13. With prior generations of language models, fine-tuning was absolutely necessary to get LLMs to perform adequately at specialized tasks. But as LLMs have grown larger and more capable, fine-tuning has generated diminishing returns. From ChatGPT-4 onward, how much fine-tuning can improve model performance at specialized legal tasks is questionable.

118 Zhao et al., *supra* note 20, at 23; Dawson, *supra* note 113, at 30.

119 Dawson, *supra* note 113, at 30.

120 *Id.* at 30–31.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

proceeds with the highest-rated option.[121] As a result of RLHF, the LLM will be encouraged to select responses that more closely conform to users' preferences.[122]

Retrieval-augmented generation—referred to as "RAG"—is a clever method for connecting LLMs with information beyond their training data.[123] Before an LLM system that uses RAG responds to a prompt, the system retrieves information that may be relevant to answering the user's query and edits the prompt to include this information. RAG is useful for grounding LLM responses in accurate information and reducing hallucinations.[124] When an LLM system needs to use knowledge beyond an LLM's training data, such as information from a specific legal case or unique information about a client, RAG is essential. But its effectiveness depends on the RAG system identifying the right information to retrieve and inject into the prompt.[125] RAG works best for factual questions with answers contained within an information database.[126] Since RAG was first introduced, many methods have been developed to assist both in correctly identifying the relevant information to retrieve and in organizing databases for optimal retrieval.[127]

Function calling and tool use allow an LLM to interact with external systems to do things like retrieve information from search engines, perform calculations with calculator software, or input data into databases.[128] If an LLM system has been programmed to call

121 RLHF was a huge factor behind the ChatGPT breakthrough. It transformed language models from software that would riff on text and produce nonsense into helpful assistants able to execute tasks and carry out commands. Tianyu Wu et al., *A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development*, 10 IEEE/CAA J. Automatica Sinica 1122, 1122 (2023); Linyao Yang et al., Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling 4 (Jan. 30, 2024) (unpublished manuscript), https://arxiv.org/abs/2306.11489.

122 Mazur & Thimmesch, *supra* note 111, at 13.

123 Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in* Advances in Neural Information Processing Systems 1–2 (2020).

124 Amnon Catav et al., *RAG Makes LLMs Better and Equal*, Pinecone (Jan. 16, 2024), https://www.pinecone.io/blog/rag-study/.

125 For a more detailed examination of these challenges with retrieving relevant legal documents, see generally Daniel Locke & Guido Zuccon, Case Law Retrieval: Problems, Methods, Challenges and Evaluations in the Last 20 Years (Feb. 15, 2022) (unpublished manuscript), https://arxiv.org/abs/2202.07209; Paul D. Callister, *Generative AI Large Language Models and Researching the Law*, 53 Brief (ABA) 18, 19–24 (Spring 2024).

126 *See* Doyle & Tucker, *supra* note 19, at 196.

127 *See generally* Chris Pang, *Techniques for Improving the Effectiveness of RAG Systems*, NVIDIA (Mar. 2024), https://www.nvidia.com/en-us/on-demand/session/gtc24-dlit63076/; Yifei Wang et al., Unveiling Factual Recall Behaviors of Large Language Models Through Knowledge Neurons (Aug. 6, 2024) (unpublished manuscript), https://arxiv.org/abs/2408.03247; Xavier Daull et al., Complex QA and Language Models Hybrid Architectures, Survey (Apr. 7, 2023) (unpublished manuscript), https://arxiv.org/abs/2302.09051; Luyu Gao et al., Precise Zero-Shot Dense Retrieval Without Relevance Labels (Dec. 20, 2022) (unpublished manuscript), https://arxiv.org/abs/2212.10496.

128 Minghao Li et al., API-Bank: A Benchmark for Tool-Augmented LLMs 1 (Oct. 25, 2023) (unpublished manuscript), https://arxiv.org/abs/2304.08244; Sander Schulhoff et al., The Prompt Report: A Systematic Survey of Prompting Techniques 5–15 (Dec. 30, 2024) (unpublished manuscript), https://arxiv.org/abs/2406.06608.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

functions or use tools, the LLM, when generating a response to a prompt, is not limited to producing text—it can also generate commands for calling these functions or tools.

A wave of contemporary research focuses on improving LLMs' performance on complex tasks through novel prompting techniques. These techniques can encompass both prompt engineering and prompt architecture. Prompt engineering refers to techniques for writing individual prompts to elicit LLM responses that exhibit stronger performance at natural language tasks. Prompt architecture refers to methods for structuring multiple prompts and responses to better accomplish complex tasks through a series of LLM calls. Unlike fine-tuning or RLHF, prompting techniques are cheap and accessible.[129] Prompt engineering happens both at the user level and behind the scenes. Developers write "system prompts," which are visible to the LLM but not to the user, that provide additional instructions to the LLM, such as "rely only on the retrieved sources for information."[130]

Regardless of how software is built up around the LLM, the underlying question is still whether LLMs can do legal reasoning work. Although the software built around the LLM influences how well the application performs, the threshold question of whether LLMs can perform these legal reasoning tasks at all remains.

B. A "Jagged Frontier" of Legal Reasoning Capabilities

Somehow, becoming very good at predicting the next word in a sequence has enabled LLMs to competently perform a wide range of tasks they weren't explicitly trained to perform. LLMs can draft human-like text, translate between languages, do arithmetic, summarize text, answer trivia questions, solve riddles, classify text by sentiment or content, write and debug code, and extract information from unstructured data.[131] Often, these emergent capabilities were not present in smaller language models but unexpectedly

129 Shubham Vatsal et al., Can GPT Improve the State of Prior Authorization via Guideline Based Automated Question Answering? 1 (Feb. 28, 2024) (unpublished manuscript), https://arxiv.org/abs/2402.18419.

130 Prompting techniques such as chain of thought (CoT), tree of thoughts, ReAct, and self-consistency with CoT have been shown to improve LLM performance on complex reasoning tasks. *See generally* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in* Advances in Neural Information Processing Systems 35 (2022), https://arxiv.org/abs/2201.11903; Shunyu Yao et al., ReAct: Synergizing Reasoning and Acting in Language Models (Mar. 10, 2023) (unpublished manuscript), https://arxiv.org/abs/2210.03629; Xuezhi Wang et al., Self-Consistency Improves Chain of Thought Reasoning in Language Models (Mar. 7, 2023) (unpublished manuscript), https://arxiv.org/abs/2203.11171. Chain-of-Thought (CoT) is a prompt engineering technique for guiding an LLM to develop its reasoning step-by-step before reaching an ultimate conclusion. Wei et al., *supra*, at 1. ToT builds upon CoT with the prompt architecture of a thought tree with multiple reasoning branches. Shunyu Yao et al., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models, in* Advances in Neural Information Processing Systems 1 (2023), https://arxiv.org/abs/2305.10601.

131 Lai et al., *supra* note 71, at 188; Homoki & Ződi, *supra* note 105, at 446.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

appeared in larger ones purely due to the increased size of the models and training data, not because the LLM was trained with the goal of developing these abilities.[132]

Researchers have described LLMs' capabilities as a "jagged technological frontier" in which LLMs can consistently perform some tasks but fail at other tasks that seem to be of similar difficulty or even easier.[133] The frontier is jagged because anticipating when LLMs will succeed or fail is hard. An LLM might be able to write a sophisticated sonnet that obeys all the rules of the form but fail to write a grocery list that includes the required ingredients from a list of meals. A language model's success in one context is not indicative of its success in other contexts, particularly with reasoning tasks.[134] This makes it difficult to extrapolate general abilities from successful performances. How language models became able to perform any kind of complex reasoning tasks remains a mystery. One of the surprise breakthroughs of ChatGPT was that given enough training data and the right deep learning architecture, next-word-predictors could use language patterns to successfully answer questions that would seem to require reasoning beyond semantic context and connection.[135] But language models' success at reasoning tasks is fragile.[136] LLMs are prone to hallucinate information, cannot accurately gauge the uncertainty of their own answers, and have limited context windows that inhibit their ability to work with larger documents.[137]

LLMs struggle with tasks that require structured reasoning, and legal reasoning—despite what its critics might say—requires structured reasoning. Math is the clearest example of this weakness. LLMs' performance on math problems is inconsistent because they don't generate answers from applying mathematical concepts; instead, they rely on language patterns from their training data.[138] This means that LLMs generate answers based upon the prompt's similarity to familiar examples. When an LLM correctly answers "What is 12 times 12?" it succeeds because it has seen this equation repeatedly, not because

132 Wei et al., *supra* note 9, at 1; Jingfeng Yang et al., Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond 11 (Apr. 27, 2023) (unpublished manuscript), https://arxiv.org/abs/2304.13712.

133 Dell'Acqua et al., *supra* note 17, at 1.

134 Tyler A. Chang & Benjamin K. Bergen, Language Model Behavior: A Comprehensive Survey 18 (Aug. 26, 2023) (unpublished manuscript), https://arxiv.org/abs/2303.11504; Jie Huang & Kevin Chen-Chuan Chang, Towards Reasoning in Large Language Models: A Survey 7 (May 26, 2023) (unpublished manuscript), https://arxiv.org/abs/2212.10403.

135 Surden, *supra* note 8, at 1954; Coan & Surden, *supra* note 46, at 12.

136 Chang & Bergen, *supra* note 134, at 17; Mahowald et al., *supra* note 88, at 11.

137 *See* Banerjee et al., *supra* note 20, at 8; Zhao et al., *supra* note 20, at 19; Xie, *supra* note 20, at 2; Sherry Yang et al., Foundation Models for Decision Making: Problems, Methods, and Opportunities 17–18 (Mar. 7, 2023) (unpublished manuscript), https://arxiv.org/abs/2303.04129. We've all read the stories of lawyers who relied on ChatGPT to draft their court filings only for the judge to discover that the legal arguments cited to nonexistent (but superficially plausible) authority that ChatGPT had invented out of whole cloth. Fagan, *supra* note 13, at 49.

138 Mahowald et al., *supra* note 88, at 11; Mirzadeh et al., *supra* note 21, at 2.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

it performs multiplication. When a language model correctly answers a math problem, that's because its training data provided it with enough similar examples that language alone led the model to the right answer.[139] After observing a language model's success at answering double-digit multiplication problems, one might mistakenly conclude that the language model had learned math—or at least learned how to multiply. But the model's performance was highly dependent on the frequency of similar patterns within its training data.[140]

LLMs' proficiency with legal reasoning follows the same "jagged frontier" observed with LLM performance on other complex tasks. LLMs excel at language manipulation tasks and at legal reasoning tasks that are well represented in language patterns in their training data. But they can struggle with tasks that require abstract reasoning and conceptual thinking beyond pattern recognition. Some empirical legal research has found that, for performing certain legal tasks, LLMs are nearly as good as or even better than law students and lawyers.[141] Current LLMs can pass the bar exam, outperform law students on certain law school exams,[142] and perform slightly worse than expert tax lawyers at answering tax questions.[143] But not all of the empirical literature has found LLMs to be legal reasoning powerhouses. When presented with statutory language and asked to answer straightforward legal questions, LLMs regularly made mistakes with simple reasoning tasks.[144] These problems became even more pronounced when the LLM was provided with hypothetical statutes guaranteed not to be part of the model's training data.[145]

139 A more subtle example of LLMs struggling is their work on planning tasks. Kambhampati et al., *supra* note 88. Planning involves more than the application of a language pattern and requires reasoning through the consequences of accomplishing steps in different sequences.

140 As LLMs have become more sophisticated and scaled up in training data, the frequency of mathematical errors has decreased but not disappeared. A striking example of this phenomenon is research that tests LLMs' performance on slightly reworded math problems. Mirzadeh et al., *supra* note 21, at 1. When numerical values are altered or irrelevant details are added to the word problem, LLM performance can drop dramatically (up to 65%), even when the altered information doesn't change the reasoning steps required to solve the problem. For example, researchers inserted the italicized text into this word problem: "Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, *but five of them were a bit smaller than average*. How many kiwis does Oliver have?" *Id.* at 10.

141 Rohan Padhye, *Software Engineering Methods for AI-Driven Deductive Legal Reasoning, in* 2024 Onward! 2024: Proceedings of the 2024 ACM SIGPLAN Int'l Symp. on New Ideas, New Paradigms, & Reflections on Programming & Software 85, https://dl.acm.org/doi/10.1145/3689492.3690050.

142 Choi & Schwarcz, *supra* note 11, at 1.

143 Nay, *supra* note 12, at 1.

144 Andrew Blair-Stanek et al., *Can GPT-3 Perform Statutory Reasoning?*, ICAIL '23: Proc. of the Nineteenth Int'l Conf. on A.I. & L. (2023), https://dl.acm.org/doi/10.1145/3594536.3595163.

145 LLMs struggle to handle basic text tasks done in legal practice, but fine-tuning a model can lead to near perfect performance on these tasks. Andrew Blair-Stanek et al., *BLT: Can Large Language Models Handle Basic Legal Text?*, 216, 216 Nat. Legal Language Processing Workshop (2024), https://aclanthology.org/2024.nllp-1.18/.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

Skeptics doubt whether LLMs can be said to have an actual legal reasoning skill.[146] Their accomplishments on the bar exam and law school exams are impressive, but it's not clear what this research shows us about LLMs' legal reasoning skills in general. After observing a language model's success at performing a specific legal reasoning task, one might mistakenly conclude that the language model has learned a particular method of legal reasoning. But a model's success in one context is an unreliable indicator of its success in another context. Rather than having portable "legal reasoning" skills, an LLM's success may be limited to particular contexts where pattern recognition suffices.

Current empirical research offers an incomplete picture of LLMs' legal reasoning abilities. Most empirical research has focused on topics such as LLMs analyzing Supreme Court cases or taking the bar exam.[147] Those tests are useful and convenient as evaluative tools because they're objective or easily verifiable. But their value for evaluating what LLMs can accomplish in the practice of law is limited because lawyers don't spend their days answering multiple choice questions and arguing the hottest topics before the Supreme Court. More empirical research is needed to investigate LLM performance at day-to-day legal reasoning tasks, like those of legal aid attorneys.[148] Because LLM performance is highly contingent on the prevalence of similar examples within the model's training data, it may be that LLMs excel at mimicking reasoning patterns from famous Supreme Court cases but stumble with state-level issues that receive less internet attention.

In research aimed at filling that gap in knowledge, Aaron Tucker and I investigated LLMs' structured legal reasoning capabilities across different state-level legal domains.[149] Our experiments tested how LLMs handle legal questions when given relevant instructions from legal practice guides.[150] We designed the questions to have straightforward fact patterns that any competent attorney would be expected to answer correctly if given the same materials.[151] But the questions concerned areas of state law likely to be underrepresented in the LLMs' training data.[152]

146 *See* Engstrom, *supra* note 25, at 1008 ("Even the most cutting-edge tools struggle with language's subtleties and law's complexities and so are not up to the legal cognitions that seasoned lawyers possess, particularly the construction of mappings of doctrine."); *see also* Sarah B. Lawsky, *Computational Law and Epistemic Trespassing*, 2 J. Cross-disciplinary Rsch. Computational L. 3 (2024) (noting how "even if one can, on some level, comprehend the meaning of a legal rule in the most literal sense, understanding the rule includes understanding how and why it came to be constituted").

147 Spaić & Jovanović, *supra* note 8, at 398 (collecting examples).

148 Although Colleen V. Chien and Miriam Kim have conducted empirical research evaluating how access to LLMs can improve legal aid attorney performance. *See generally* Chien & Kim, *supra* note 34.

149 Doyle & Tucker, *supra* note 19, at 194.

150 *Id.* at 194–95.

151 *Id.* at 202.

152 *Id.* at 195.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

The findings revealed that LLMs often fail to correctly answer straightforward legal questions—even when given a practice guide with clear instructions on how to apply the law.[153] We also found puzzling differences when breaking questions into sub-questions for the model to answer one at a time.[154] Sometimes this improved performance, and sometimes it made things worse.[155] When asking a legal question piece by piece led to wrong answers, it was typically because the LLM found legal issues where there weren't any to find. Vice versa, when asking a legal question in whole led to wrong answers, it was usually because the LLM overlooked actual legal issues.[156] A vivid example comes from an experiment based on New Jersey laws concerning pretrial detention. In New Jersey, if someone is arrested for a crime, a judge is not allowed to keep them in jail by setting a money bail amount they cannot afford to pay.[157] Instead, to keep them in jail, a judge must make an order of pretrial incarceration based on their risk of dangerousness or flight.[158] We gave OpenAI's GPT-4o model a hypothetical fact pattern in which a New Jersey judge set an unaffordable money bail amount to keep a person in jail pretrial and asked GPT-4o whether this was lawful.[159] We included in the prompt an excerpt from a legal practice guide that describes the legal requirements for incarcerating someone pretrial in New Jersey. In answering this question, GPT-4o overlooked the judge's imposition of unaffordable money bail as an issue and found that the court's actions were lawful.[160] But when we split the question into two questions, one that asked whether there was a proper legal justification and another that asked whether proper legal procedures were followed, GPT-4o correctly identified that the imposition of unaffordable money bail was unlawful.[161] If we asked these same questions to a human lawyer, we'd expect them to answer correctly regardless of whether the question was broken up into parts.[162]

One might expect that an LLM would make consistent legal reasoning errors based on a particular legal misconception the LLM has: The LLM would consistently give incorrect answers to questions that present a specific legal issue. One might also expect an LLM to handle routine legal tasks easily but sometimes make mistakes regarding subtle or

153 *Id.* at 198–202.

154 *Id.* at 202–03.

155 *Id.* at 202.

156 *Id.* at 202–03.

157 Robert Ramsey, 31 N.J. Practice Series, Criminal Practice and Procedure § 19:1 (2024).

158 *Id.*

159 Doyle & Tucker, *supra* note 19, at 197–98.

160 *Id.* at 201.

161 *Id.* This doesn't mean that breaking up a legal question into sub-questions is always the right strategy. With other legal questions, breaking up the question into sub-questions resulted in worse performance because it resulted in the LLM finding legal issues where there weren't any to find.

162 The fact patterns were written to be legally unambiguous questions that are clearly answered by the practice guide excerpts included in the prompts, so 100% accuracy should be achievable.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

ambiguous points: The LLM would reliably produce correct results for clear-cut cases but struggle with cases that are in a legal gray area and are similarly challenging for humans. What we observed was different. The LLMs' errors were not related to consistent legal issues and were not related to more subtle or ambiguous points. Rather, the errors were both inconsistent and unpredictable.

Before conducting these experiments, our assumption was that LLM performance would improve with the addition of the practice guide and would improve more when the legal questions were broken down. It didn't. The LLMs produced inconsistent results when asked to apply unfamiliar legal rules to facts, even when given precise instructions.[163] More research is needed on the topic, but these findings suggest that LLMs do not have a reliable "general" legal reasoning ability that allows them to apply law to facts in novel legal and factual contexts.[164] LLMs' surprisingly strong performance in other contexts may be due to the prevalence of past examples of these kinds of questions within the LLM training data.

LLMs' underlying architecture can help to explain why their performance at legal reasoning tasks is so jagged. Language models accomplish complex tasks by relying upon language patterns that are contextually similar to the language of the prompt, not by applying an abstract logic that can be applied from context to context. The pattern-matching nature of LLMs explains their facility with tasks that rely heavily on linguistic convention and form. They excel at summarizing documents, drafting routine correspondence, and generating text that mimics familiar styles and formats.[165] Legal reasoning tasks do involve the application of language patterns. There's a reason that the CRuPAC or IRAC structure is drilled into the mind of every 1L.[166] But answering legal questions often requires more structured thinking and logical analysis than pattern application alone.

LLMs produce text that looks like the product of human thought—but they don't arrive at those outputs by thinking like a human.[167] The text they generate is a reproduction

163 *Id.* at 194. Across different legal contexts, no single approach consistently outperformed the other approaches, and no single language model consistently outperformed the other models. *Id.*

164 "Instead, in order to make significant progress, a concentrated and domain-specific effort may be required that is specifically designed to promote the capabilities of language models to engage in forms of legal reasoning." Frankenreiter & Nyarko, *supra* note 18, at 72.

165 Homoki & Ződi, *supra* note 105, at 444–45; Bogdan Padiu et al., *To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review*, 15 Info. 662, 671 (2024).

166 Tracy Turner, *Finding Consensus in Legal Writing Discourse Regarding Organizational Structure: A Review and Analysis of the Use of IRAC and Its Progenies*, 9 Legal Commc'n & Rhetoric 351, 352 (2012).

167 Although language models produce human-sounding language, their relationship to language is far from human. When we talk, our speech reflects our thoughts and beliefs about the world. We think thoughts—both in language and outside language—that isn't captured in the language we write down or speak aloud. Language models don't have thoughts or maintain consistent beliefs. Bender et al., *supra* note 108, at 7; Mahowald et al., *supra* note 88, at 1.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

of patterns of words combining and recombining in context with one another. The core mechanism of a language model is to generate a response by predicting, one token at a time, what text is most likely to follow. It is not oversimplifying things to say that LLMs are next-word prediction machines.[168] It seems fantastical that the text that an LLM produces is, at bottom, just the result of a math equation, but it is. To an LLM, the words it processes and generates are not even words; they're numbers that represent words.[169] Yet somehow, by discerning statistical relationships between these numbers-as-words across vast datasets of examples of human language, LLMs are able to generate new text that humans find meaningful.[170]

An LLM doesn't reason through problems but rather generates text based on similarly worded problems in its training data. This can lead to the peculiar (and now familiar) phenomenon of LLMs generating text that superficially resembles legal analysis but is also hollow, a pastiche of relevant terms masquerading as rigorous thought.[171] The bottleneck on what LLMs can accomplish is the nature of the patterns they are trained to detect and

168 Korinek, *supra* note 22, at 4.

169 Arbel & Hoffman, *supra* note 105, at 25.

170 Most of the time.

171 Legal hypotheticals offer a useful illustration. When a language model is asked to write a legal hypothetical, the hypothetical it produces will conform to Socratic dialogue hypotheticals at a surface level. But the hypothetical will often fail to conform with the purpose of a Socratic hypothetical: to challenge a listener's understanding or definition of a legal rule by revealing a contradiction, a recognized exception, an absurd or unfair result of applying the rule, or something else about the rule that may be problematic. Elizabeth Garrett, *The Socratic Method*, Univ. Chi. L. Sch. (1998), https://www.law.uchicago.edu/socratic-method. In my experience, the more obscure the legal rule, the more likely the LLM will produce a weak hypothetical. With legal rules that are part of the 1L law school curriculum, an LLM has many example hypotheticals in its training data that it can regurgitate more or less verbatim.

A language model's legal hypotheticals bear only a surface-level resemblance to the hypotheticals a law professor might ask a student or an appellate judge might ask an attorney at oral argument because semantic patterns fail to capture the connection between the initial legal rule and the hypothetical application of that rule. While discussions of the critical thinking challenges of the Socratic method tend to focus on the students on the receiving end of legal hypotheticals—and rightly so!—the creation of a legal hypothetical requires critical thinking as well. The process of writing a legal hypothetical can be broken down into at least two steps. In the first step, one must imagine how an articulated rule will apply across a range of hypothetical contexts with the goal of finding unintuitive or unfair results, contradictions with other rules or underlying principles behind the rule, or some other problematic result. In the second step, one must select a particularly challenging example from among the hypotheticals and present it back to the listener to reconcile with their definition of the legal rule.

The semantic pattern of Socratic dialogue may be that of a general rule and a specific application with an unwanted outcome or bad result. The legal hypotheticals that language models generate seem to follow this pattern well enough. But the semantic pattern fails to unearth how the Socratic dialogue is not so much about identifying bad outcomes as it is about exposing contradictions. A good legal hypothetical will identify not just an unfortunate outcome but a scenario in which the outcome of the legal rule contradicts the values that justify the rule. This connection is latent in the text of the general rule and the legal hypothetical. Therefore, it can be overlooked by a language model that generates text based on language patterns alone.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

replicate: language patterns.[172] Rather than having a general legal reasoning ability, LLMs have pockets of success and failure that seem highly dependent on prevalence of similar examples within the training data.[173] LLMs are a tremendous breakthrough. But just because they can produce human-like outputs, that does not mean they can *always* produce outputs that are equivalent to human cognition.[174] Even if we can't precisely anticipate when an LLM will succeed or fail at a particular legal reasoning task, we can understand that limits on LLMs' general legal reasoning capabilities result from LLMs' underlying logic being a logic of language patterns.[175] The jagged frontier of LLMs' legal reasoning capabilities is inherent in the way LLMs are constructed.

III. PROSPECTS FOR AUTOMATED LEGAL REASONING

How does the LLM breakthrough affect the prospects for developing software that autonomously performs legal reasoning tasks? Can automated self-help tools be developed without programmers having to painstakingly translate the law into hard-coded instructions for a computer to follow?

172 This also suggests that improving automated legal reasoning may require more than just larger models with more training data. Problems that cannot be decomposed into or framed as language patterns will remain beyond an LLM system's purview.

173 It's worth noting that this capability is not a function of the complexity of the task. LLMs are able to perform some highly complex legal reasoning tasks while failing at straightforward legal reasoning tasks. Accordingly, complexity is a poor proxy for anticipating the success of an LLM's performance at a task.

174 Nor is there reason to assume that this obstacle can be overcome by simply training LLMs on more data. Additional training data is unlikely to be an adequate solution to this problem. Reliance on training data would require exhaustive testing for each corner of the law before reliable performance can be assured for that legal domain. LLM systems would be too slow to react to changes in case law and legislation. Reliance on training data would also result in poor LLM performance in sparse domains without many examples from the case law.

175 On a related note, we need stronger mental models for how LLMs work. Whether we realize it or not, we use mental models with technology all the time. Mental models are frameworks for expectations of how a technology works. Laura Rook, *Mental Models: A Robust Definition*, 20 The Learning Org. 38, 40 (2013). A good mental model doesn't require the user to understand the technology at a technical level but serves as a reliable framework for expectations when using that technology. For new technology, mental models often start by analogy to what is already known and are gradually built up over time. At one point, the mental model of a car was a horseless carriage. That's gone by the wayside as our experience with cars has grown, but most drivers today still have no idea how a car works. Nonetheless, these drivers have reliable mental models for operating a car and can anticipate what a car is capable and incapable of doing, what to expect under different conditions, and what feedback from the car would indicate a serious problem. At present, we're like the moviegoers in the apocryphal story of those who attended one of early cinema's first film screenings running screaming from the theater upon seeing a train barreling straight at the audience. The story is amusing and quaint to modern sensibilities. But to those moviegoers, no life experience had involved seeing a train hurtling toward them without the threat of its running over them. Likewise, before LLMs arrived on the scene, no life experience had involved seeing swaths of cogent text not written by a human intelligence.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

This part advances two arguments. First, the prospects for legal automation reflect the jaggedness of LLMs' reasoning abilities. For legal domains and fact patterns within the jagged frontier of LLMs' capabilities, automation becomes feasible. Tasks beyond that frontier remain out of reach. Second, because LLM performance is unreliable and unpredictable, legal automation with LLMs will require extensive testing and iterative redesign to ensure that the software performs well on the tasks that are within its capabilities.

A. A Jagged Frontier of Automation

Because LLMs' legal reasoning capabilities form a jagged frontier, the frontier of potential automation is also jagged. In other words, LLMs' uneven performance at legal reasoning will give rise to uneven possibilities for automation. LLMs are competent at performing some legal reasoning tasks and incompetent at performing other legal reasoning tasks.[176] Tasks within the jagged frontier of LLM capabilities are ripe for automation, while tasks at the border of or beyond the jagged frontier are not.[177] Even if LLMs have the ability to perform only a limited set of legal reasoning tasks—constrained by both the legal subject area and the nature of the task—that set of legal reasoning tasks can be automated. Given LLMs' strengths and weaknesses, you wouldn't want to rely on one as your lawyer. But you might want LLMs to do *some* lawyering.

Automating legal reasoning tasks might seem to require an AI that possesses a general legal reasoning skill, a transferrable analytic ability that works across contexts.[178] And there's a huge gap between LLMs' capabilities and a lawyer's legal reasoning capabilities. You cannot simply hand an LLM a set of legal rules and a fact pattern and expect the LLM to successfully apply the law to the facts—except sometimes you can. LLMs' success at performing legal reasoning tasks is uneven, not universally bad. These uneven and unpredictable results upend our conventional understanding of how legal reasoning skills are supposed to operate. While lawyers might exhibit varying levels of proficiency at legal reasoning tasks, a person's performance at legal reasoning tasks is much more consistent than the erratic performance of LLMs. A lawyer's expertise may vary from one context to another—you wouldn't expect a tax lawyer to know the details of employment law

176 Neel Guha et al., LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models 4 (Aug. 20, 2023) (unpublished manuscript), https://arxiv.org/abs/2308.11462; Weikang Yuan et al., Can Large Language Models Grasp Legal Theories? Enhance Legal Reasoning with Insights from Multi-Agent Collaboration 1 (Oct. 3, 2024) (unpublished manuscript), https://arxiv.org/abs/2410.02507.

177 For legal tasks that cannot be automated, AI can still augment lawyers' efforts by retrieving, organizing, and presenting relevant information and simulating legal arguments, hypotheticals, and questions. Augmentation—as opposed to automation—is worth further study and may have a greater impact on legal practice in the foreseeable future than attempts at automation. *See* Chien & Kim, *supra* note 34, at 908–10.

178 *See* Frankenreiter & Nyarko, *supra* note 18, at 72 ("To be successful [at performing legal reasoning], legal tech applications would need to automatically extract from relevant texts a structured representation of legal concepts and their interconnections (what we refer to as a 'legal ontology').").

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

litigation—but a lawyer's underlying reasoning skills ought to remain consistent across domains. Inform that tax lawyer of the relevant legal rules for employment litigation, and she can apply them to the facts of a case. But LLMs' underlying reasoning skills are not consistent across domains. Give an LLM an eminent domain fact pattern and the legal rules from federal constitutional law, and the LLM can write up a competent legal argument. Hand the same model the same facts and the legal rules from Minnesota's state constitution, and it might produce nonsense.[179] LLMs' jagged performance defies our expectations. A lawyer who can make persuasive arguments under federal eminent-domain law can manage the state-law version too. The lawyer might need to do their research first, but you'd expect them to be able to construct a coherent argument—and you wouldn't expect them to draw illogical conclusions and invent fake precedents. LLMs demonstrate no such reliability or transferability of legal reasoning skills,[180] leading some to the conclusion that the current generation of AI will not prove capable of automating legal reasoning work.[181]

But LLMs' peculiarities could also lead to a different conclusion: Our conception of automation is inapt given how LLMs work. LLMs lack a general reasoning ability, and their outputs can be unreliable, but neither of these shortcomings is as formidable of an obstacle to automating legal reasoning work as it may appear.[182] Automation of legal reasoning may seem to require machines that possess the same "general" legal reasoning skill that human lawyers possess, but the history of automation suggests otherwise. The idea that a general legal reasoning ability must precede automation reflects our profession's limited experience with automation. The legal field has been slow to mechanize because legal reasoning is so steeped in human language, and—until the recent LLM breakthroughs—computers have been unable to comprehend, process, or work with human languages.[183] Legal work has largely resisted automation, leaving the profession with little experience of how automation typically progresses.

179 The LLM might make up precedent, misunderstand the doctrine, or write an argument that defies logic entirely. *See* Doyle & Tucker, *supra* note 19, at 200 (documenting this phenomenon with Minnesota eminent-domain law).

180 Their outputs may mimic the language of legal reasoning, but they don't arrive at those outputs by doing legal reasoning.

181 *See* Frankenreiter & Nyarko, *supra* note 18, at 72.

182 These arguments are rooted in the practical possibilities and limits of current technology, not a potential sci-fi future that tends to dominate legal scholarship discussions of AI and legal reasoning. *See* Engstrom, *supra* note 25, at 1003 ("[D]ebate has tended toward gauzy and futurist meditations about a legal system populated by 'robo-judges' and 'robo-lawyers,' or even an eventual state of 'legal singularity,' when machines perfectly predict the outcome of every case before it is filed.") (citing Eugene Volokh, *Chief Justice Robots*, 68 Duke L.J. 1135, 1137–42 (2019); Milan Markovic, *Rise of the Robot Lawyers?* 61 Ariz. L. Rev. 325, 349–50 (2019); Benjamin Alarie, *The Path of the Law: Towards Legal Singularity*, 66 U. Toronto L.J. 443 (2016)).

183 Legal reasoning work has become computerized. That's vastly different than legal reasoning work becoming mechanized.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

Automation happens piecemeal. It rarely begins with machines that possess general human abilities. Instead, machines first master narrow tasks within broader human domains. Parts of manual tasks become partially mechanized long before machines rival humans at performing an entire range of tasks.[184] Machines may lack human flexibility and adaptability, but they can perform specific functions more efficiently, cheaply, and tirelessly than humans.[185] Consider textile production during the Industrial Revolution. Machines like the Jacquard loom didn't possess the general skills of artisanal weavers.[186] Instead, they performed narrow tasks efficiently at scale.[187]

Or take the modern example of self-driving cars. At present, self-driving cars cannot match human drivers' general ability to drive.[188] Like LLMs with legal reasoning, they have a jagged frontier of capabilities. Self-driving cars depend heavily on training data and on mapped information about specific roadways, and they struggle with situations human drivers are expected to handle, such as driving in snow or heavy rain.[189] As a result, we do not have ubiquitous self-driving cars but rather pockets of automation in sunny cities with modern infrastructure, like Phoenix and San Francisco, where self-driving cars have been extensively tested before deployment.[190] A jagged frontier of capabilities has resulted in pockets of real automation based on the machine's pockets of competency.

The same pattern can apply to LLMs and legal reasoning. LLMs excel at some tasks and fail at other tasks that would seem to be equally simple and straightforward. The jagged frontier of LLM abilities creates a corresponding jagged frontier of potential automation. A key challenge to automation is demarcating this frontier, as its contours are unknown. This can be done only through extensive, iterative testing and design.

184 Discussions of LLMs' capabilities are often focused on whether LLMs represent or portend a breakthrough in what the computer science community has termed "general intelligence." General intelligence is a type of intelligence that can "possess a reasonable degree of self-understanding and autonomous self-control, and ha[ve] the ability to solve a variety of complex problems in a variety of contexts, and learn to solve new problems that [it] didn't know about at the time of their creation." Hugo de Garis, *Artificial Brains, in* Artificial General Intelligence vi. (Ben Goertzel & Cassio Pennachin eds., 2007). It's an interesting topic, but a preoccupation with the prospect of an AI with general intelligence can minimize or obscure the impact that AI with narrow intelligence can have.

185 Hamming, *supra* note 23, at 19 ("It has rarely proved practical to produce exactly the same product by machines as we produced by hand. Indeed, one of the major items in the conversion from hand to machine production is the imaginative redesign of an equivalent product.").

186 *The Story of the Jacquard Loom*, Sci. & Indus. Museum (June 25, 2019), https://www.scienceandindustrymuseum.org.uk/objects-and-stories/jacquard-loom.

187 *Jacquard Loom*, Age of Revolution, https://ageofrevolution.org/200-object/jacquard-loom/ (last visited Aug. 19, 2025).

188 Maya Indira Ganesh, *The Ironies of Autonomy*, 7 Hums. & Soc. Scis. Commc'ns, Nov. 2020, at 1, 4.

189 *Id.*

190 Matthias Harlwig, Self-Driving and Cooperative Cars 51 (2024).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

B. *Iterative Testing and Design*

The greatest hurdle confronting automating legal reasoning tasks is the unpredictability and unreliability of the results that LLMs produce. The cornerstone of trust for any legal AI process is portability. Portability refers to a process's ability to perform well in a variety of contexts.[191] With language models and legal reasoning, this would refer to an application's ability to produce reliable results across legal and factual contexts. Unfortunately, a language model's success in one context tells us remarkably little about its likely performance on even closely related tasks. The reasons for success or failure lie buried within the black box of an LLM's internal parameters, which makes it impossible to extrapolate generalizable conclusions from limited examples of an LLM's success or failure.[192] An LLM's success performing one task tells us little about how successful it will be performing another task, even when the two tasks look similar in difficulty. The only way to know whether an LLM can handle a given legal task is to test its performance. If the LLM fails, the next steps are to iteratively redesign the software processes around the LLM and keep testing.

The challenges with building LLM-powered software that can autonomously perform a legal reasoning task are that one cannot know in advance if the software can perform the task, if the software can perform the task consistently across a variety of possible inputs, or if the software is prone to hallucinate or make other catastrophic errors. For any legal task in any legal domain, reliable performance can be determined only as the result of extensive testing related exclusively to that task and domain. Reliable automation of legal reasoning tasks is not an automatic consequence of LLM capabilities but rather requires careful software design. The current impediment to automation is not so much that the underlying technology hasn't progressed enough; it is that the underlying technology has not been fully harnessed and tested within legal domains.[193]

Compared to traditional computational law approaches, automation with LLMs requires less up-front work but more backend validation. Traditional "computational law" systems were too brittle and arduous to build and maintain.[194] These systems

191 Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency 59, 61, https://dl.acm.org/doi/10.1145/3287560.3287598.

192 That said, we can make some educated guesses based on how LLMs have been trained. Given that modern LLMs train on essentially the entire internet, they likely perform better on legal reasoning tasks that are well-represented on websites, blog posts, social media, and digital publications. Conversely, LLMs are likely to struggle with niche, obscure, or unconventional legal reasoning tasks that no one bothers to blog about. But this is a rule of thumb for relative likelihood of success of different tasks rather than an absolute assurance of successful performance.

193 This is not unique to law. Many fields *might* be subject to automation in large or small part but share this bottleneck.

194 Eboigbe, *supra* note 4, at 7.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

demanded Herculean up-front labor: Experts had to formalize the law by turning every rule, exception, and factual permutation into code before the program could run.[195] LLMs escape these constraints because they're flexible and will attempt to perform any task, not just the tasks that they've been trained to do (or are competent to perform).

LLMs' flexibility is both an asset and a liability. Prototypes are easy to assemble: Just prompt the model with legal rules and facts, and the LLM will produce something that looks a lot like legal analysis. The surprise breakthrough of LLMs is that models trained to perform one task—predict the next word in a sequence—were able to do other tasks they weren't trained to perform, such as take the bar exam, revise computer code, or develop a strategy for negotiating a raise at work.[196] This is what makes generative AI special, particularly in comparison to traditional software that cannot perform any tasks out of context. LLMs overcome the "brittle" nature of logic-based systems for legal reasoning. There's no need to settle upon a theory of legal reasoning and hard code every rule and exception. An LLM can pick up nuances from training data and prompt examples, handle open-textured provisions, and incorporate diverse sources of knowledge.

But this flexibility comes at a cost, and the bitter must be taken with the sweet. An LLM's ability to perform out of context makes evaluating its reliability incredibly difficult.[197] With traditional software methods, evaluating reliability is not just possible, it's baked into development as a required part of building the software in the first place. But when an LLM is deployed in a new context, there's often no test data available to use to evaluate the model's performance.[198]

Understanding and calibrating LLM performance requires substantial backend work: testing the system's performance across countless hypotheticals, designing guardrails to block nonsensical or dangerous outputs, and iteratively tweaking prompt design to improve overall performance.[199] With generative AI, the drudgery of formal modeling is exchanged for the drudgery of quality assurance. Testing methods can be used to verify LLM-powered software's capabilities in general and the authenticity of particular outputs. Test suites are a ubiquitous practice in software development to ensure the reliability of computer code. A test suite is a collection of tests that can be automatically and

195 Coan & Surden, *supra* note 46, at 9.

196 In machine learning parlance, language models are described as being terrific few-shot and zero-shot learners, meaning that they can perform new tasks without having been intentionally trained to perform them. Jason Wei et al., Finetuned Language Models Are Zero-Shot Learners 1 (Feb. 8, 2022) (unpublished manuscript), https://arxiv.org/abs/2109.01652.

197 Li et al., *supra* note 128, at 1.

198 Difficulty in evaluating AI for legal tasks can also vary based on the clarity of correct labels and observability of relevant features. Sayash Kapoor et al., Promises and Pitfalls of Artificial Intelligence for Legal Applications 1 (Jan. 10, 2024) (unpublished manuscript), https://arxiv.org/abs/2402.01656.

199 *Id.* Another challenge is that some tasks have clear, correct answers, while others don't.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

periodically run on a software program to ensure that the program is functioning correctly.[200] To ensure LLM reliability at performing a legal reasoning task, a test suite for this software should include tests that verify the validity of the LLM components' responses to a comprehensive set of inputs that are representative of the types of inputs the LLM system may receive.[201] These tests can be run before the software is deployed in a production environment, while the software is deployed, and whenever the software is modified to ensure that the LLM components are still functioning correctly.

Beyond testing whether an LLM can perform a legal reasoning task, software developers can iteratively redesign other software components around the LLM and test its impact on the LLM's performance.[202] As described in part II.A., language models can be modified, and software can be built around the model to do things like automatically provide the LLM with particular information, evaluate the LLM's outputs, and chain sequences of LLM calls and responses together. Contemporary LLM-powered software typically includes both an LLM and a hard-coded infrastructure around the LLM to constrain its output, to enable it to use tools such as a calculator or web search, and to direct it to act in a certain way or to consider particular information when generating responses.[203]

And what about hallucinations? Hallucinations are a serious problem, but they've become a bogeyman in legal circles.[204] LLMs are unreliable, but with proper design, it's often possible to build reliable systems from unreliable components.[205] The prevalence and

200 Test suites are not designed to generate any inferences about LLMs' general abilities. They are here only to evaluate LLM performance at a particular task.

201 Lawyers will need to be involved in the creation of these test suites because they involve domain-specific legal questions. Transparency in the contents of these tests and the software's performance on the tests can help assuage justified fears of errors and hallucinations.

202 *See* Homoki & Ződi, *supra* note 105, at 438; Wu et al., *supra* note 121, at 10; Yang et al., *supra* note 121, at 4; Lewis et al., supra note 123, at 1–2.

203 It's hard to know what capabilities might be unlocked using techniques like fine-tuning and RLHF for improving LLM performance. And it is challenging to identify the limits of those improvements. When working with LLMs, we can articulate their tendencies, but we can't confidently articulate their limits. Whatever the process produces is necessarily a combination of an LLM's underlying capabilities, the prompts that a person has designed, and the architecture of the application itself. When the application seems unable to accomplish a human reasoning task effectively, it's difficult to pinpoint whether that limit is because the task is beyond what the application can accomplish because of underlying deficiencies with LLMs or if that reasoning task could be done more effectively with different prompting or different system architecture that segments or structures the task differently. As a result of this, when an LLM system continues to fall short at some task despite our best efforts at varying the prompting and architecture, all that we can conclude is that its tendency is to fall short. In many instances, we can't document this as an insurmountable limit.

204 There are now many of examples of attorneys and pro se litigants filing AI-authored documents that cited to nonexistent cases as governing legal authority. Fagan, *supra* note 13, at 49. But these stories are as much about user error and the danger of improperly trusting unvetted technology as they are about the danger of LLM hallucinations.

205 The computer science adage "garbage in, garbage out" is not always true. *See* Hamming, *supra* note 23, at 246–50.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

impact of hallucinations and errors can be mitigated by methods that include constraining the tasks an LLM is asked to perform, developing robust testing methods, and incorporating other language models to verify the accuracy of LLM outputs.[206] Preventing legal AI software from making obvious, egregious errors—such as citing to fake cases—is straightforward. The software can include validation routines that do things like checking case citations in the software's output against a database of legitimate cases and flagging any anomalies. But subtler errors might be harder to address, particularly the kinds of errors that can't be easily and consistently identified by automated routines. Total error elimination is impossible. But that's true of human systems too. The viability of an automated legal AI software is ultimately a policy question of error tolerance. What kinds of errors can be tolerated, and at what rate?[207]

LLMs lack a general "legal reasoning" ability that transfers across contexts. What they offer instead are narrower—but potent—pockets of competency. LLMs open pathways for automation within domains where they can perform reliably. The challenge lies in identifying these domains and optimizing performance through systematic evaluations and iterative redesign.

IV. AUTOMATING ACCESS TO JUSTICE

This part more closely examines how the prospects for automated legal reasoning affect the prospects for automating access to justice. The generative AI breakthrough may result in greater access to justice, but it's not guaranteed. The fundamental tension between the optimist and pessimist views of technology and inequality remains the same. The optimists hope that technology will democratize access to resources, while the pessimists fear that technology will exacerbate unequal access to resources. Part III's assessment of the prospects for legal automation suggests that investment choices will determine whether automation narrows or widens the justice gap. More people will receive the legal services they need if investments are made to develop LLM-driven software targeted at the legal issues of low-income litigants. But disproportionate investment in proprietary AI-driven tools that benefit high-income litigants risks exacerbating inequality. The jagged frontier of legal automation will become more jagged based upon where investments are made along that frontier. Principles of transparency and open-source research must guide the development of AI tools for legal aid to prevent the proliferation of untested tools that harm the people they're meant to serve, to foster the development of shared standards

206 Automating some legal reasoning tasks may not require achieving the impossible goal of eradicating hallucinations at the level of individual LLM responses. The overall system may be able to achieve this goal instead.

207 Part IV.B addresses this question in greater depth.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

for evaluating AI, and to inform the public and potential litigants of the risks and benefits of these tools.

A. *Investment and Inequality*

Let's define the automated legal aid self-help tool we're contemplating: Software that takes as input a user's text description of the facts of their legal case and returns text that explains the legal implications of those facts.[208] The software relies on LLMs—rather than a manually coded, step-by-step process of transforming facts into legal conclusions—to discern these legal implications. The simplest—and most unreliable—version of an automated self-help legal aid application would be a website with an LLM chatbot that has access to relevant legal rules and existing self-help documents. A user could ask the chatbot a question; the system would try to provide the LLM with excerpts from legal rules and self-help documents that are relevant to that question; and the LLM would provide an answer. More complex software could include multiple LLMs in conversation with one another, separate processing of user questions or LLM responses to refine or filter inputs and outputs, or traditional logic-based workflows for processing different kinds of questions.[209] However this software is designed, it's essentially software that performs a multistep process to respond to a user's query, with the software using an LLM as one component of the process rather than having an LLM *be* the entire process.

Developing effective automated self-help tools depends on LLMs having a baseline competency with the job they're asked to perform. This requires mapping the jagged frontier of LLM abilities for legal aid problems. Creating this domain-specific and task-specific map will require deliberate investment aimed squarely at the legal domains and issues of low-income litigants. The current barrier is not so much that the underlying technology hasn't progressed enough to make automation possible but rather the friction of testing and wrangling the technology for specific legal domains.[210]

208 Given legitimate concerns about software violating rules of professional conduct or practicing without a license, it may be important to limit what the application provides to the user as potentially relevant information rather than legal advice. *See* Norton, *supra* note 27, at 195; Chien & Kim, *supra* note 34, at 920. Regardless of how far into the practice of law the software might intrude, the application would still need to correctly perform some legal reasoning steps to be able to provide a user with the right information to consider.

209 "Agentic AI" refers to LLM systems in which an LLM is able to direct the course of its own actions. Nay, *supra* note 12, at 14–15. It often involves a central "planning" LLM developing a plan that is then broken into steps for other LLM agents to execute. Neurosymbolic systems are those that are a hybrid of LLMs and traditional logic-based systems for performing reasoning tasks. Meng Fang et al., *Large Language Models Are Neurosymbolic Reasoners*, Proceedings 2024 AAAI Conference on Artificial Intelligence 17985, 17985, https://ojs.aaai.org/index.php/AAAI/article/view/29754.

210 As shown in part III, it takes time and work to map the jagged frontier, push its boundaries, design software built around LLMs, and test their performance. And each self-help application requires its own domain-specific testing and software design.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

The hype surrounding generative AI can make it seem like self-help software should be able to be conjured up in a session or two of vibe coding. And with generative AI, it's easy to build LLM-powered prototypes that will produce some results. But these results are prone to serious errors such as misinformation, hallucinations, and biased responses. With only a few lines of code, one can build an LLM system that spits out answers to legal questions. But turning that prototype into a production-ready application still demands time and money. Useful self-help tools won't automatically spring forth as a downstream consequence of LLMs' improved general capabilities. The difficulty is in using the testing and design process to build up prototypes into production-ready applications. Breakthroughs in generative AI can't change the fact that improving access to justice requires the serious investment of resources, time, and expertise to assess LLM capabilities and operationalize them in practical self-help software.

AI may not eliminate investment as a prerequisite for access to justice, but AI may promise a greater return on investment than traditional approaches to building self-help software. LLM-powered software might achieve better results at lower costs and enable solutions in domains where traditional rule-based systems proved infeasible. Technology also attracts funding that traditional legal aid doesn't.[211] The excitement surrounding AI solutions might draw investments that never would have materialized for attorney salaries.[212] Some funders may prefer up-front technological investment over ongoing personnel costs, though all technology requires some level of maintenance and support.

The jagged frontier of automation will become even more jagged as investment flows to different legal domains. In some areas, investment in AI tools for access to justice is necessary to prevent the justice gap from widening. Automation is most likely to occur with highly revenue-generating processes because those processes promise the greatest return on investment. But there's not much money to be made from creating legal aid AI tools. As Colleen Chien has noted, "[w]hen an innovation need is unique to a poor or underrepresented population, there is a greater risk of it going unmet as innovative energy is drawn elsewhere."[213]

AI threatens to widen existing disparities in "David and Goliath" legal proceedings—debt collection, immigration cases, and eviction proceedings—that pit a low-income litigant against a well-funded adversary.[214] These sophisticated, repeat players have every

211 *See* Rediet Abebe, Designing Algorithms for Social Good, 192–95 (Dec. 2019) (Ph.D. dissertation, Cornell University), https://doi.org/10.7298/n8w3-8629 (demonstrating how technology can draw attention to social issues that would otherwise be overlooked).

212 For example, the Legal Services Corporation has allocated funds through technology initiative grants to encourage the development of AI in legal aid programs. Norton, *supra* note 27, at 194.

213 Colleen V. Chien, *The Inequalities of Innovation*, 72 EMORY L. J. 1, 16 (2022).

214 Hassan Kanu, *Artificial Intelligence Poised to Hinder, Not Help, Access to Justice*, REUTERS (Apr. 25, 2023), https://www.reuters.com/legal/transactional/artificial-intelligence-poised-hinder-not-help-access-justice-2023-04-25/.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

incentive to invest in bespoke AI tools, and many already have.[215] There is an inexorable logic and momentum to incorporating AI given how inexpensive and efficient it is compared to human labor. Machines are cheaper, indefatigable, and consistent.[216] Higher-income litigants have an incentive to stockpile data, build ever-more-effective AI tools, and keep these gains private.[217] Without commensurate investment for low-income litigants, the justice gap will only widen as their powerful adversaries continue to accumulate an advantage.

At the same time, we shouldn't assume that generative AI will offer the same benefits to indigent and well-heeled litigants. It may be that generative AI can provide more help to pro se litigants than it can to licensed attorneys. Some experiments have found that access to LLM technology substantially improves the performance of lower-achieving workers while not benefitting top performers much at all.[218] This effect has also been found in experiments testing law students' exam performance with and without AI assistance,[219] suggesting that AI might be "an equalizing force in a notoriously unequal profession."[220]

For legal domains in which the parties stand on roughly equal footing, such as family law, AI-driven tools might reduce inequality and increase the baseline level of legal support for all parties involved. Rather than have two unrepresented parties stumble through a legal proceeding that neither is equipped to handle, automated self-help tools can help both parties understand their rights and formulate arguments more effectively. If both sides can tap into low-cost, well-tested self-help AI tools, the price of competent guidance may plummet, narrowing rather than widening the gap.[221]

The effects of automation on our court systems should not be overlooked. Yonathan Arbel has noted that an unintended effect of automated legal services may be inefficiency and backlogs.[222] AI's potential to reduce access-to-justice barriers suggests a possible litigation boom as more individuals can engage with the legal system effectively.[223] While AI can improve access to justice, it may simultaneously threaten judicial economy. The

215 Engstrom & Gelbach, *supra* note 5, at 1039.

216 Thomas G. Gunn, *The Mechanization of Design and Manufacturing*, 247 Sci. Am. 114, 15 (1982), https://www.jstor.org/stable/24966684.

217 And the closed nature of legal AI start-ups and major players like Westlaw and LexisNexis limits the ability to even assess the current state of investment in legal AI and what it has produced.

218 Dell'Acqua et al., *supra* note 17, at 11.

219 Choi & Schwarcz, *supra* note 11, at 6 ("The worst-performing students benefited enormously from AI, with gains of approximately 45 percentile points. By contrast, the best-performing students received worse grades when given access to AI, experiencing declines of approximately 20 percentile points.").

220 *Id.*

221 And when one party has a lawyer and the other party does not, these self-help tools can mitigate the disadvantage the unrepresented party faces.

222 Arbel, *supra* note 10, at 566.

223 *Id.*

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

increased use of AI tools could lead to a rise in the volume and verbosity of caseloads, straining the balance of judicial resources against the demand for legal services.[224] Historically, challenges to judicial economy have been met by adjusting procedural rules and substantive doctrines, often resulting in the shrinking of substantive rights.[225] This could mean that litigants who gain access to justice might encounter narrower rights and stricter administrative procedures.

B. Transparency and Standards

Transparency of self-help legal aid software is necessary to protect the public. Proliferation of hastily built, low-quality, automated self-help tools for legal aid is a serious risk. Unlike traditional legal technology and software, generative AI prototypes for a legal self-help tool can be constructed in minutes. All a programmer needs to do is connect an LLM to a database of legal rules and self-help-guide information scraped from existing PDFs, place it on a website, and watch the AI tool generate answers to any question a user provides. But without rigorous testing and redesign, the system will hallucinate case law, misinform users, and provide bad legal advice. These tools need rigorous evaluation and careful design. Testing and development should be transparent and open source so the public and researchers can know what is being adopted and hold developers accountable.

LLM systems offer greater flexibility than traditional logic-based approaches. This flexibility allows them to handle a broader range of inputs, thereby imposing less stringent requirements on litigants to frame their issues within strict legal frameworks before even using self-help materials. But flexibility comes at a cost: Errors cannot be completely eliminated. Even the best generative AI systems will produce unanticipated errors precisely because their responses have not been preprogrammed.

It would be a mistake to reject LLM-powered tools because perfect accuracy is unattainable. The status quo produces plenty of errors. When litigants can't receive help or can't navigate the resources available to them, they're prone to making significant errors.[226] Automated self-help tools for legal aid raise a question common to other automated tools, from self-driving cars to pretrial risk assessment algorithms: Against what standard should we be judging this technology?[227] A mediocre chatbot may be prone to error, but is it still better than the status quo? The question isn't whether to accept errors at all but what types of errors can be tolerated, at what rate, and in what domains. This is a fundamentally normative and domain-specific question. The analysis must consider not

224 *Id.*

225 *Id.* at 549.

226 Norton, *supra* note 27, at 211.

227 *See* Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,"* 80 FED. PROB. 38, 38 (2016) (providing an example of this debate in the pretrial risk assessment context).

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

just error rates but consequences—a system should permit fewer major errors than minor ones. Different legal contexts will demand different standards. No uniform standard will likely emerge, but transparent discussion about error prevalence and mitigation techniques is necessary in a democratic society. Litigants deserve to know a chatbot's error rate before they rely on it. Developers who publish those figures invite critique, but they also build public trust.

The benefits of open research extend beyond the oversight of individual tools. Developing these systems transparently, with shared research, enables the creation of collective standards for testing and evaluation. At present, basic principles for evaluation and testing of legal AI tools can be articulated, but the precise methods can't be prescribed in advance—they need to be discovered in the process of building the software. Open dialogue among developers across jurisdictions can help clarify the optimal evaluation methods. When researchers share their code, others can build upon and refine that work. A set of shared methods and metrics can emerge only from developers sharing their work with one another. With open-source software development, standard practices for testing AI self-help tools may emerge organically.[228] Without that free exchange of ideas, a multiplicity of standards will be developed independently, in private. And the only standard against which to judge the fairness and effectiveness of these tools will be the developer's assurances that the tool works as intended. Most research into generative AI and law is currently being done behind closed doors by companies that look to profit from the adoption of this software. As software is rolled out, the processes that LLMs follow are hidden from users and outside researchers. Without open-source research, the legal community, policymakers, and the public won't know how to evaluate the reliability and fairness of these tools, how to assess the risks, or even how to identify the right questions to ask.

CONCLUSION

For better or worse, America has turned to technology to solve its access-to-justice crisis. Generative AI, and LLMs in particular, open up new opportunities for legal automation, potentially delivering low-income litigants legal information and advice tailored to their circumstances. For decades, getting computers to perform legal reasoning has been a futile endeavor because traditional software is too brittle and inflexible to accommodate the messy, humanistic work of legal reasoning. LLMs challenge our expectations of computers' strengths and weaknesses. Somehow, becoming very good at predicting the next word has

228 This includes both local standards for similar legal domains and more global standards across the legal field itself. As standard practices emerge, the costs of building these systems could be reduced even further because future self-help systems could adopt existing best practices rather than requiring legal experts and programmers to rediscover and reinvent these practices from scratch.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025

enabled LLMs to competently perform a wide range of tasks they were never trained to perform. With complex tasks like legal reasoning, LLMs' capabilities form a jagged frontier: They excel at some tasks but fail at others. Tasks within the jagged frontier of LLM capabilities are ripe for automation, while tasks at the border of or beyond the jagged frontier are not. Because an LLM system's success performing one legal reasoning task is not indicative of success performing other tasks, automation requires domain-specific and task-specific investment to refine prototypes into production-ready applications through careful testing and design.

AI's impact on access to justice is not predetermined; it is dependent upon targeted research and investment. Because LLM abilities are not portable across legal domains and legal tasks, developing an automated system for any task in any legal domain requires extensive evaluation and redesign related exclusively to that task and domain. The jagged frontier of automation will become even more jagged based on the parts of the frontier that receive this investment. Only with targeted investment for developing automated legal aid tools and a commitment to transparency and open-source software development can technology live up to its potential for bridging the justice gap.

Downloaded from http://direct.mit.edu/ajle/article-pdf/doi/10.1162/AJLE.a.6/2567948/ajle.a.6.pdf by guest on 23 December 2025